MOLECULES, MICROBES, MINDS, AND MACHINES:

TOWARDS A SCIENCE OF THE SUBJECTIVE

Matthew M. Hurley

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Cognitive Science Program,
Indiana University,
September 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee:

_____

Douglas R. Hofstadter, PhD

_____

Colin F. Allen, PhD

_____

Randall D. Beer, PhD

_____

Hamid R. Ekbia, PhD

August 29th, 2018

For Justina and Colette

Acknowledgments

Countless generous people have helped carry me through the process of writing this dissertation. It would not have been possible without them. Some of these people were willing to entertain my ideas; some argued fervently against them; and some showed me altogether new ways of thinking. Among these people there are a handful whose contributions were quite a bit larger than others, and who deserve special mention now.

Above all, I must thank my wife Justina Fan. She has given me unending inspiration, faith, patience, and support, year after year. She believed in the seeds I dreamed of planting and in my ability to germinate them. And she gave me the earth, the water, and the sun to help them grow. Justina is not a scientist, but she has surely proven to be a patron of the sciences, and the world would be a better place if it had more people who share her values.

I am also indebted to the committee of advisors who provided me with the delicate balance between the freedom and encouragement necessary to explore these coarsely charted waters and just enough guidance to avoid sinking my ship.

It was Hamid Ekbia's (2008) book *Artificial Dreams* that provoked me to look more deeply into the topic of teleology in the first place. Hamid made it clearer to me than anyone else had, that if we are to one day understand ourselves well enough that we might build machines that are "like us", we will first have to reckon with the fact that we want these machines to derive their purposiveness from their status as *organisms* rather than as *artifacts*. I daresay this project might never have been conceived if it were not for Hamid, and I appreciate his guiding hand along the way.

Without Randall Beer I could neither have begun my project nor completed it. In the courses he taught, Randy introduced me to the richness of taking the dynamical-systems perspective and he thereby forever transformed my understanding of the world. The ideas in this work don't

appear to be directly based in dynamical-systems mathematics, but in actuality they were deeply influenced by Randy's teaching of that way of thinking. I only found out later that it also was Randy who advocated my acceptance to the Cognitive Science Program in Bloomington in the year that I applied. I am immensely grateful for being given that opportunity.

Colin Allen has given me years of both gentle critique and intentionally adversarial inspiration. In philosophy, that is the best kind of friend one can have. And if your own personal devil's advocate is as sharp as Colin is, then your choice is either to be stubborn in your mistakes or to move your position more towards that of the devil's advocate. On every topic in philosophy, Colin has more to teach than I will ever be able to absorb. He was kind enough to invite me to his dissertation-writer's reading group every semester. And he is responsible both for moderating and for informing my positions on countless topics. He is also responsible for introducing me to most of the important nuances of the philosophical literature on functions, and for trying to help me to more clearly understand the writers from that tradition.

I owe an extraordinary dose of gratitude to Douglas Hofstadter, the chair of my committee, who has been as much a patron of this project as my wife has. Doug has had the kindness and patience to give me a space in his Fluid Analogies Research Group for many years now. And he has supported me throughout that period despite my consistent failure, one expedition after another, to bring back any compelling maps, tantalizing treasures, or even a reasonably complete ship's log. Graduate students rarely find a principal advisor willing to give the kind of open-ended freedom and encouragement that Doug provided—especially to a student whose ideas are only loosely and tangentially connected to the foci of his own research program. At least as valuable as Doug's nurturing manner was his intellectual influence. As will be apparent to anyone familiar with his work (or any reader of this dissertation), many of Doug's own ideas have profoundly influenced

both the core elements of my thesis and the adjunct analyses that scaffold various facets of the topic. I couldn't have found a more fitting advisor to chair my committee.

Although the four professors mentioned above formed my official committee in Bloomington, Daniel Dennett from Tufts University also took on an informal advisory role on this project. Dan offered himself as a hesitant-yet-hopeful sounding board for a number of these ideas. He regularly suggested literature that might help me find new ways to frame things. And his occasional reassurances helped give me some confidence that, whether my hypothesis turns out to be right or wrong, there is value in producing it. As a previous advisor, collaborator, and friend, Dan has so deeply influenced me as a thinker that I often find my ideas framed in terms I feel he might use, and I appreciate many of the perspectives he has taught me to take over the years. I also owe Dan my deepest gratitude for introducing me first to Doug's work, and then to Doug.

I have also had the pleasure of being assisted by a handful of other outstanding colleagues. Alexander Ince-Cushman and Eric Nichols deserve significant thanks for the many hours they each have spent helping to work over some versions of these ideas with a degree of seriousness that those early versions didn't obviously deserve. Eran Agmon, Nicholas Bishop, David Bender, Seth Frey, Deran Garabedian, Jonathan Hurley, Tim Lucas, and Paul Queior have all read and commented usefully on early drafts of chapters. And I have had meaningful and influential conversations or correspondence with, or have been helped in my work in other ways by Raffaello D'Andrea, Mark Bedau, Asaf Beasley, Howard Berg, Robert Bowers, Emel Gencer, Chris Harshaw, Helga Keller, Brent Kievit-Kylar, ChiaHua Lin, Robert Mahaney, Ruth Millikan, Gerardo Ortiz, Susan Palmer, Robert Rose, Alejandra Rossi, Leo Trottier, and Jason Yoder. Thank you all.

Matthew M. Hurley

August 25th 2018

Preface

Although most of us—myself included—tend to treat science as the jurisdiction of experimentalists, there are in reality many phases to the scientific treatment of a topic, and many varieties of evidence that can be brought to bear in developing a richer understanding of the facets of the world that interest us. What you will find in this work should properly be considered an informed hypothesis—a work of *theoretical* science at the intersection of philosophy, biology, physics, and cognitive science.

The hypothesis I offer is far from the first word on these topics, and even further from being the last word. But it is a hypothesis that I believe has its place near an important inflection point within the ages-old discourse about biological and human nature, as well as the physical world that serves as their context.

As of this moment, that hypothesis comes with hundreds of small incompletions and dozens of larger ones. Critics who head into it mining for weaknesses will no doubt be rewarded—they will find more ore than they know how to smelt, and I wish them an enjoyable expedition. But despite all its problems, the value in producing a hypothesis of this sort is not in getting it entirely right, nor in successfully defending it against all the early criticisms that might be leveled against it. It is in stimulating the thoughts of a new generation of scientists to begin to explore these topics from a new perspective. I hope that, if I accomplish anything here at all, it is that.

Matthew M. Hurley

MOLECULES, MICROBES, MINDS, AND MACHINES:
TOWARDS A SCIENCE OF THE SUBJECTIVE

In this work I pull together many long-explored ideas on agency, vitality, function, and goal-directedness in an attempt to explain how the *subjective* properties of our world arise from its *objective* constituents. The main ingredients of my theory are: (1) a distinction between *comparative* and *evaluative* norms, aimed at dividing the philosophical notion of normativity into two separate problems; (2) a view of life and vitality as a form of resistance to *material* disorder (in contrast to Schrödinger, who saw them as a form of resistance to *energetic* disorder); (3) the idea that although no organizational pattern in the world has an intrinsic *function*, certain organizational patterns in the world do possess intrinsic *goal-directedness*; (4) a new mathematical characterization of the metaphysical notions of identity and value; (5) a set of distinctions, based on my new view of identity and value, that allows different kinds of orderliness in the world to be classified; and finally and most importantly, (6) a theory of teleology, rooted in all these ideas, which I believe can underpin an eventual science of the subjective.

Table of Contents

List of Figures

Chapter I

Purpose: An Introduction to Teleology

Biology has not so far been able to analyse and paraphrase in precise and exact scientific, i.e. physico-mathematical, terms, the really fundamental characteristics of observed life in the full and original meaning of that word: that apparent purposiveness and goal-directed character of vital activities which make the distinction between living and non-living systems in nature one of the most important distinctions Man draws among the objects of his environment: those characteristics of observed life which invite us to think of living things as somehow endowed with a soul, as somehow capable of personal relationship; those characteristics which invite us to think of living nature as permeated by intelligence and purpose, and whose loss we mourn when death occurs.

– Gerd Sommerhoff (1950:1)

The earth is a rare and strange place. It is energetic and it is dynamic; teeming with life and pulsating with industry; seething and swarming with activity and objects that we don't know to exist anywhere else in the universe: not only organisms, but also the products they produce, the husks they leave behind, the societies they comprise, and the cultures they sustain. In all but the most remote corners, our planet is saturated; it is steeped and it is sodden . . . it is dripping with purposeful items and events. And yet, in comparison, broad swathes of the rest of the universe appear to be barren of this activity and devoid of these objects. The problem that Sommerhoff approached, some sixty-five years ago, goes unexplained still: Science has not been able to analyze, in any precise terms, *the fundamental characteristics of observed life*—the purposiveness and goal-directed character of vital activities that make the distinction between living and non-living systems one of

the most important distinctions we can draw. The goal of the current work is to characterize that distinction.

It turns out, however, to be a rather multifaceted distinction. One might suggest that the core of this work is the issue of *subjectivity* or *agency* since it relates to the way subjects or agents come to exist in an objective world, to the way a thing (or a pattern[1]) can transcend the objective nature of its physical constitution and come to have a *perspective*. One might also say—indeed the title of this chapter seems to say—that the focus is what's called *teleology*, also known as *purposiveness* or *goal-directedness*. It is a theory of how some items in our world might be said to be the kind of thing that is *for* (or that exists *in order to* do) something or other, the kind of thing that serves a function or that behaves toward the attainment of goals, the kind of thing that is purposeful.[2] The eminent French biologist Jacques Monod termed this notion *projectivity* to highlight the fact that, in some sense, organisms transcend objectivity not only through their perspectival subjectivity but moreover through their *behavior*—organisms set themselves incessantly to working on *projects* (1971). They *do* something for some *reason*. Nothing else in the world comes even close to being so industrious.

Following Monod's observation that the class of projective agents maps onto biological organisms, one might also say that the core topic here is *life* itself. It is a biological thesis . . . one that centers not on how the molecular, genetic, physiological, and ecological processes of life work, nor on the evolutionary details of how the diverse forms of life have come to be, but more specifically on *what accounts for their vitality*.

While this may look like a patchwork quilt of concepts, if you read on I think you'll eventually come to conclude, as I have, that the picture is not really pieced together from patches; it is a tapestry woven from fibers and the concepts that make up those fibers are so intricately

[1] The term "thing" is obviously not well defined and may lead to interpretation that is influenced by an object-chauvinism that takes material boundaries too seriously. I will discuss my use of the alternative term "pattern" in

[2] As it turns out, this focus on purposiveness depends also on a deep understanding of *existence*, and so, the theory is also one of *ontology*—a theory of the kinds of things that can exist in our universe.

interlaced that theories of them could not easily be separable into individual theoretical strands. The project of explaining the phenomena described above is, I believe, an all-or-nothing affair (see also Deacon's, 2013, treatment of many of the same phenomena, which he groups together under the label "ententional", though I'll steer clear of using that term).[3]

Some parts of what is at stake here are obvious. A definition of vitality—or of life, taken independently of cell-biological constraints such as nucleic acids, ribosomes, protein synthesis, and lipid bilayers—will help us understand what kinds of things could possibly be agents in the universe, what we should watch out for when searching for life, and what we should strive for when trying to foster it[4]. It also will give us a basis for making ethical judgments about how we would like to treat various patterns that we may encounter in our world. A theory of goal-directedness and of the ways that goal-directed things achieve their goals will help us understand not only the relationships between agents, their bodies, and their behaviors, but also the relationships between agents and their artifacts, between ourselves and our technologies, and between ourselves and the artificial intelligences that some scientists hope one day to build.

Other parts of what is at stake may be less obvious at first, but will become clearer in time. The more focal topics above are just as intricately interwoven with, and inseparable from, the subjects of *identity* and *value*, as well as *autonomy*, *cognition* and *will*. A close look at the first pair (in chapters II, IX and X) will provide the footing for our central topics to build upon, while the theory

[3] This all-or-nothing attitude also reflects my affinity for the Duhem-Quine thesis of confirmation holism, which claims that scientific statements cannot meaningfully be tested absent the context of heaps of other related theoretical claims. Empirical work reflects upon the predictive power not of an individual narrow hypothesis, but of a complex, intertwined network of related theoretical pieces. And because of this, theories should be understood as being (allegedly) explanatory not in isolation, but in relation only to broader theoretical frameworks (Duhem 1954; Quine 1951). Narrow scientific hypotheses may be simpler to work with but, no matter how clearly a set of experiments might seem to answer them, they are more likely to be misleading precisely because they are disconnected from a broader context of relationships.

[4] In, say, artificial life and artificial intelligence endeavors as well as in NASA's and SETI's search for (signs of) life beyond earth.

I'll advocate for those central topics (in chapters VIII through XI) will have broad implications for the latter trio.

Currently, the issues in the collection of italicized terms scattered across the past few pages are all topics of philosophy. A primary goal of the current work is to nudge them all closer to being topics in the realm of science. In 1944, Erwin Schrödinger noticed, much as Sommerhoff did, the vast difference in our abilities to explain the behaviors of animate and inanimate matter. Schrödinger suggested that there should be laws—laws of purpose (or of subjectivity, or of life), just as elegant as many of the laws of physics—that can provide a foundation for a quantitative study of life and cognition (Schrödinger 1944; see also Bohr 1933; T. Nagel 2012; Oparin 1964; but for reductionist dissent, see *e.g.* Schaffner 1967 and also, despite his observing the projectivity of organisms, Monod 1971). Nowadays the idea that such laws might exist seems to be viewed with general disregard but I hope this work will be part of the path toward showing that, in this regard, Schrödinger was right[5].

[5] Schrödinger (1944) also suggested that the laws to be discovered would center on what he called the "aperiodic crystal"—the chromosome—that at the time had yet to be illuminated as being what we now call DNA (Watson and Crick 1953). Kauffman (1997) argues against this aspect of Schrödinger's view of life. On Kauffman's view it is not that DNA does not play a role in life, but only that the behavior of DNA (and also of natural selection) is not the central phenomenon producing vital behavior. In fact, he offers a compelling explanation of how self-organizing autocatalytic sets of molecules can be lifelike in the absence of genetic polymers. I'll have much more to say about Kauffman's theory later.

A. Goals

It is the all-pervasive presence of this apparent purposiveness in life processes, and the resulting possibility of thinking of living systems in terms of the 'goals' towards which their activities are directed, which is mainly responsible for the radical difference between the ways we think about living and non-living things.

—Gerd Sommerhoff (1950:5)

The notion of goals or of being goal-directed is one of the central themes in this work and, whether it was by chance or by necessity, it happens to be the germ from which my exploration of the rest grew. Thus it seems appropriate (to me) to begin by characterizing goals and goal-directedness both in terms of our basic intuitions about the subject and in terms of the most widely agreed-upon philosophical reflections on the subject. These topics will all be revisited in more depth, in later chapters. Let's begin with ends.

Ends

The concept of *goals* is often taken to be synonymous with *aims*, *ends*, *objectives, intentions,* and *purposes*, among other things. It seems to be the *result* that is important to goals . . . at least certainly more so than the *means,* which can vary widely. But are ends truly what goals are? And in any event, what are ends?

Certainly no moment in *time* is privileged or exceptional in any way. We could say that the "end" of the day is at midnight, but that choice was arbitrary and historical. It was defined out of

convenience and could as well have been sunset, or sunrise, or even "bedtime"[6]. Few other objects or events have any reference to points in time in their specification, but those that do are also chosen by discrimination of time's relevance to our interests. A race, though timed, comes to an end when the judges deem that the finish line has been crossed regardless of when it happens or how long it takes.[7]

It is more common to consider an end to be a *state* in a physical system. For instance, one might want to say that a marble coming to rest at the bottom of a half-spherical bowl or a pendulum bob coming to rest directly below its pivot point is an end. But what is actually ending when such rest is achieved? In most cases, the pendulum bob or marble are still vibrating to some small degree—being repeatedly perturbed away from the rest position by some forces and then coming back to (something that resembles) rest. In all cases, the bob or the marble is hastily circling the center of the earth at up to perhaps a thousand miles an hour[8] and the earth itself is zipping around the sun at an absurd rate, itself dwarfed by the nearly half million miles per hour at which the marble, earth and sun all together make their way around the galactic core of the Milky Way. The determination of states such as "being at rest" requires interest-dependent (that is to say, subjective) judgments. Their definitions are formal prescriptions—they are based, firstly, upon boundaries where we choose to divide a continuous variable (such as a rate of motion) into discrete categories and, secondly, upon a frame of reference with respect to which the variable can be measured for comparison against the threshold (for instance, rate of motion with respect to the room, but not with respect to the moon or the sun).

[6] Or it could have been based instead on the apparent sidereal day, which, though in some sense more "universal", is a far less convenient schedule to live by, since sunshine has biological consequences.
[7] In a similar vein, the painter Rembrandt is often quoted as having remarked, "A painting is finished when the artist says it is finished."
[8] Depending upon its latitude.

An alternative description of the pendulum or the marble-in-bowl system can be given in terms of dynamical systems theory—the mathematics that describes the qualitative behaviors of physical systems in terms of the differential equations representing the forces that govern the flow of those systems. In either of our scenarios, the marble or the pendulum bob is moving in the flow of the dynamics created by a number of factors, including the gravitational field, the damping friction of air resistance, and the other molecular forces of the various parts of the pendulum or of the bowl and the marble (*e.g.* friction, tension, torsion, and so on). As a result of all these factors, there exists what dynamical systems theorists call an "attractor". An attractor is a mathematical object that represents the eventual position of items in the flow of the system. A point near the bottom of the bowl is an attractor for the marble, and a point below the pivot is an attractor for the pendulum bob. In other cases attractors are not points but instead circles or ellipses (for instance, in the case of a stable orbit) or some other more complicated or chaotic patterns (such as the now-famous double-lobed Lorenz attractor).

An interesting consequence of this view is that, even when an item is apparently "at rest", it is still behaving identically within the flow of the system as it was before it came to "rest"—it is flowing towards the attractor in the system on a pathway defined by the equations (ultimately, the *forces*) that govern the flow. This behavior—*flowing*—is never-ending unless the system itself is perturbed (the string released, the bowl tipped over, or the local gravitational environment altered) and, even then, flowing continues; it just continues towards a different attractor defined by the parameters of the new system. In other words, motion may appear to end, but flow is endless; "staying there" is just a special case of moving. What has actually ended when a pendulum or marble appears to come to rest is only *subjectively perceived* motion—motion within the range of sensitivity to which a particular perceptual system is tuned.

The conclusion I want us to draw here is that objective "ends" do not exist in our world. Flow is eternal and no state, defined in terms of either time or space, can count as an end without subjective determination of one sort or another. The hope of some modern thinkers—to disguise the long-noticed subjectivity of goals in the clothes of objective ends—cannot succeed if ends themselves are not objective.

Agency

Whether or not there can be an objective definition of ends, goals are not usually seen to be merely resultant physical states in the sense implied by the marble-in-bowl or pendulum examples[9]. Today, in the distant wake of Newton's (1687) publication of the laws of motion and universal gravitation, most scientists and even most laypeople don't believe (or even wonder whether) marbles or pendulums are goal-directed. The laws of physics sufficiently explain the behavior of these non-biological objects and so no question remains for the notion of goal-directedness to answer[10]. Physical states or attractors may be seen as *ends* (though, see above), but those ends are not, in and of themselves, *goals*.

Instead, we usually identify as a goal only a particular kind of end that has some relationship to an agent or a subject. Certainly this is the case with canonical goals—our own human goals—wherein *we* are the agents for whom certain ends become goals . . . the subjects who either intend or benefit from the attainment of those particular ends (Bedau 1992a, 1992b). The notion easily generalizes also to organisms as agents who similarly appear either to intend or at least to benefit

[9] Mayr (1974) labels phenomena of this sort "teleomatic" because by approaching attractors automatically such systems seem to be *end*-directed without being *goal*-directed.
[10] See also Newton's own "Hypothesis I" of his "Rules for Reasoning in Natural Philosophy", discussed on p. 158 of this manuscript.

from certain ends (though we'll delve into the question of intention and of the representation of ends, shortly).

So we might say that a particular person had the goal of getting a marble to stop at a particular position (say, in a game of marbles) but few if any of us (ever since Galileo, Newton, and Laplace) would say that a marble itself has the goal of reaching the bottom of the bowl or that a pendulum bob is goal-directed toward coming to rest below its pivot. It requires a subjective view, or an agent, to turn a particular end into a goal (although currently we do not know how). So not only is a subjective viewpoint required to convert states into ends but it also takes a subjective perspective to transmute ends into goals.

The fact that goals are unavoidably subjective is not a threat to an underlying objectivist view of the world, as long as the status of being a goal-directed agent is somehow defined objectively, and then subjective goals are definable in terms of such agents. If this is the case, then subjectivity arises unproblematically from the subject. The question that remains, and that lies at the heart of my project here, is: What defines a subject or an agent? In the next chapter I'll *begin* to answer that question by introducing the notions of identity and value, which, together, I call the "fundaments of subjectivity" and which I take to underlie agency. It seems to me that any theory of goal-directedness worth its salt will be based in an understanding of agency founded on precise, physically justifiable, explanations of these fundaments of subjectivity. In Part II, I will attempt to give a version of those physically justifiable explanations.

About the Future

Another thing about the notion of goals and of being goal-directed that has given no end of headaches to centuries of philosophers and biologists is the fact that these finalistic[11] concepts refer to events or states of affairs that have not yet come to pass. Goals and goal-directed behaviors are, in a sense, about the future.

The problem, of course, is that this seems to allow that future events or states (the goals) may somehow direct current behaviors, providing the paradoxical concern—called "backwards causation"—that something in the future can *cause* what is happening now or what has already happened in the past (see *e.g.*, Braithwaite 1954; Mayr 1974; Spinoza 1677). If drinking your blood causes a mosquito's proboscis to puncture your skin—if we can only explain the (earlier) puncturing in terms of the (later) feeding—then causation appears not to work only in the way that physics would have us believe.

The reaction to this seemingly unnatural, anti-causal circularity has been twofold. First, it is usually pointed out that, in the case of human goals, it is not actually the future state but the *mental representation* of the future state that causes behavior. It's the *thought* that counts. Since a representation occurs prior to behavior, it can produce that behavior through forward-causal means, and so there no longer seems to be anything "backwards" to worry about. I will argue below that this line of reasoning, while reasonable, is nonetheless oversimplifying, firstly because it ignores the nonhuman remainder of goal-directed creatures, and secondly, because the conscious, cognitive goals of humans that do explicitly represent future states are subservient to the same kind of

[11] The word "final", in its teleological sense, means "end-related" and so the derived term "finalism" is more or less synonymous with "teleology"; it implies a belief that some things are done in order to achieve an end. We'll trace the history of this terminology in Chapter III.

representation-free[12], goal-directed nature that exists in those nonhuman organisms. For instance, I may have the consciously represented goal of baking brownies this afternoon, but that ambition is subordinate to the involuntary, instinctive, and unrepresented goals of satisfying my hunger and my sweet tooth. I'll talk again about representation (and about baking brownies) below, but we can supplement the conclusion here with the observation that we humans also often perform fully automatic, involuntary goal-directed behaviors where representation plays not a subordinate role, but no role at all. For instance, we don't think about or represent digestion of our latest meal before doing it, though the act is goal-directed toward the production of usable catabolites (biological building blocks). It is clear then that representation is not as central as it may at first seem, even to human goal-directedness. Ultimately our account of goal-directedness and our elimination of backwards causation will need to be the same in the case both of psychological and of non-psychological organisms, and it will need to be given in terms agnostic of mental representations[13].

The second reaction to backwards causation that modern theorists have had is simply to discount the existence of goals and goal-directedness in non-psychological biology. The future can't cause the past and so (the thinking goes) we can only describe what happens in most of biology as being *seemingly* or *apparently* but not *truly* goal-directed (see chapter VII). The mosquito is not puncturing your skin *in order to* drink your blood; it is only puncturing your skin *and then* drinking your blood. The fat-tailed dwarf lemur is not accumulating fat deposits in its tail now *in order to* use that fat to fuel its body during hibernation (Dausmann *et al.* 2004); it is only accumulating fat in its

[12] The term "representation" is controversial in cognitive science, especially when used to claim that some living or cognitive agent *lacks* representation in some way. To clarify (but not to enter on either side of that debate), my use here of the term "representation-free" is meant to denote situations in which I believe the agent being analyzed is *not consciously aware* of any representation, whether or not representations exist in it.

[13] However, this is not to say that the goal-directedness is not represented *structurally* in some manner; it is only to say that it is not represented *to* the individual. If a suitably knowledgeable observer, armed with a sophisticated enough theory of life and cognition, were to view the structure of that individual, it might in principle be possible to "read off" the goals of an individual, regardless of whether that individual was aware of them.

tail first *and then* burning those calories during hibernation. As Pittendrigh (1958) put it similarly: the turtle is not coming ashore *in order to* lay its eggs; it is only coming ashore *and then* laying its eggs.

The problem of backwards causation is a serious one and so we must choose one of three paths. Either (1) we follow the popular modern scientific tradition and take goal-directedness to be merely apparent or (2) we awkwardly alter our theories of physics and causation and take goal-directedness to be backwardly causal (despite nothing else in the world appearing to be) or (3) we find an alternative way to conceive of goal-directedness that does not contradict physical causation. Later, I will advocate a view along the lines of this third path. Rather than eschew backwards causation, I'm going to attempt to give a logical and materialistically consistent account of it.

Reasons

The kind of ends that we call goals are often conceived of in terms of their use as a subspecies of reasons or explanations. When we think about a fly that we see bouncing against a window, we make sense of this behavior in terms of the reason the fly is doing so. The question that comes to mind most immediately is "*why*" or "*to what end* is the fly doing that?" or "*what* is the fly doing that *for*?" Trying to describe the random-seeming pattern of persistent bounces doesn't make sense except in the light of it being done for some reason. Even if we asked the seemingly more objective question "what is the fly doing?" the best answer still would be one that provided not just the literally requested description—what Dennett (2014) calls a "process narrative" (*e.g.* "bouncing repeatedly against the window")—but also a reason in terms of an end (*e.g.* "trying to get outside").

What our curiosity is really about, and what "trying to get outside" really answers, is the what-for question. We consider it an appropriate question to ask because a process narrative does

not seem to fully explain the fly's behavior[14]. Of course a process narrative can explain such systems as the marble-in-bowl because the "why" in "why does the marble do that?" only goes as far as "how-come", never leading all the way to "what-for"[15,16]. The only thing we find ourselves curious about is how come the marble goes to the bottom of the bowl, and so we can be satisfied by an answer that does not refer to a goal . . . an answer such as "gravity pulls the marble down, momentum carries it back up, and this would continue eternally if it were not for friction damping the momentum over time".

Many theorists have noted the distinction between mechanistic, how-come reasons and teleological, what-for reasons (Ayala 1970; Brandon 1981; Cummins 2002; Dennett 2014, 2017; Haig 2013; Mayr 1961; see also Plato's *Phaedo* and *Timaeus*, and Aristotle's *Physics* and *Metaphysics*), and a number of those attempting to account for aspects of teleology have framed their discussion in terms of the question "what is it for?" (*e.g.*, Bedau 1992b; Cummins 2002; Kitcher 1993; Wright 1973). However, as I will explore in chapter IV, the many possible interpretations of the term "for" in the question "what is it for?" can, if we are not careful, lead to much confusion.

Representation

On the face of it, the issue of representation is important to goal-directedness, at least in terms of archetypical, human goals. In most cases we need to think about what we are going to do before doing it (van Leeuwen 2016). Could you succeed at, or even have, the goal of buying eggs

[14] Notice that I am not saying quite *what* about its behavior the process narrative doesn't explain. That is the job for the entire dissertation, not just the introductory chapter.

[15] Some common linguistic signals that indicate we are looking at a what-for claim are when an explanation has in it the terms "for the sake of", "for the purpose of", or "in order to". Sometimes, "in order to" is even shortened simply to "to": What did you open the window for? I did it *to* get some fresh air. In any such case the explanation is being given in terms of some end that presumably explains the behavior or the existence of the item in question.

[16] Douglas Hofstadter has pointed out to me that some languages (for instance, Russian) have two different words that translate to the English term "why", the meanings of which correspond to the "what-for"–"how-come" distinction.

without having imagined buying eggs? We plan our projects, plan our days, plan our careers, and plan our retirement, and to the degree that we plan well, we achieve our goals[17]. But as I began to discuss earlier, many human goal-directed behaviors are neither mediated by nor ultimately goal-directed because of representational thought. For example, although representational reasoning usually plays some role in the process of grabbing a snack from the cupboard, it is not those ideas but rather our hunger (or perhaps a sociocultural codification based ultimately in a biological history of hunger) that drives us to snacking in the first place. Representation, rather than being *necessary* for goal-directedness, may instead be just one particularly *effective* way for an agent that is goal-directed to accomplish its goals.

Still, it is common for thinkers to differentiate representational goal-directedness and non-representational goal-directedness, and to insist that the difference is a significant one. Most writers carve up what-for reasons into two further categories. I am conflicted about employing the terms "conscious" or "cognitive" or "intentional" to sketch this boundary[18], but those are the kinds of terms that are usually used: There are some things that are done, or that are there, for the sake of some conscious, intentional, human goal and some things that are done or are there for the sake of organismal well-being. Let's look at some examples.

In the first category, a (human) behavior may have been performed for the sake of attaining some end (eggs are bought in order to bring home eggs with which to cook); or an artifact may have been created for the sake of attaining some end (a basket is made in order to be able to carry eggs home from the market); or an artifact may be employed for the sake of attaining some end (the basket is used in order to carry eggs home from the market). In all these cases, we understand these

[17] And to the degree that our goals are realistic, and with a bit of luck, and so on . . .
[18] My hesitation stems, firstly, from a reluctance to limit terms such as "cognitive" and "intentional" only to humans, and, secondly, from a sense that the term "conscious" is not well-defined.

ends as having been entertained in the mind—represented—prior to the performance that brought them about.

In the second category, a (nonhuman) behavior may have been performed for the sake of attaining some end (a nest is rearranged by a chicken in order to cushion the egg that is about to be laid); a product of an organism may have been created for the sake of attaining some end (a nest is first constructed in order to safely hold a clutch of eggs); a feature of an organism may exist for the sake of attaining some end (the oviduct and cloaca exist in order to aid in the careful release of eggs from the uterus of the hen into the nest); or an organism may go through some developmental process for the sake of attaining some end (eggs are formed in the uterus of a hen in order to reproduce). Unlike the cases in human teleology, the hen does not think about or plan the various components of her process of egg production and care[19].

André Ariew, following David Charles (1995), categorized the examples of teleology found in Aristotle's writing along lines such as these. (I should note that I don't mean to single out Ariew for any particular reason, except that he, nicely and concisely, summarizes a number of common impressions about the relationship between teleology and representation.) While Aristotle himself found the teleological similarities between humans and other living beings more compelling than the differences[20], Ariew focuses on those differences.

> [Agency-centered teleology] and [Teleology pertaining to natural organisms] are *distinct* notions of teleology: Aristotle should have used two words to distinguish them. Agent-specific teleology . . . is purposive, rational, and intentional, and

[19] Well, perhaps this is debatable in the case of rearranging the nest; we might attribute intention to the hen at some level of that process (though it is debatable as to whether that is a correct attribution). But there are plenty of other examples of 'instinctual' behaviors in nature to which we wouldn't attribute intention, especially as we move to non-vertebrate branches of the tree of life. Even in humans, the reflexive eye-blink is an unconscious and unplanned behavior that is nonetheless end-directed (toward keeping foreign objects out of the eye).

[20] "It is absurd to suppose that ends are not present [in natural organisms simply] because we do not see an agent deliberating." (*Physics* 2.8)

> represents an external evaluation. The goal is the object of the agent's *desire* or choice. . . . Teleology pertaining to natural organisms is distinct: *non*-purposive (though seemingly so), *non*-rational, *non*-intentional, and *immanent*—that is an inner principle of change. The goal is *not* an object of any agent's desire. (Ariew 2002:9)

Ariew does call both by the name "teleology", but it seems to me he doesn't really mean that, since for him teleology pertaining to nonhuman living beings (he calls them "natural organisms") is *non-purposive*. It's just *similar* to teleology but . . . well, as he says, "distinct".

I can see why many people take these to be distinct notions, but I am inclined to agree with Aristotle: While there *is* a distinction, I don't think it is a *teleologically* relevant one. The two categories are similarly directed towards ends, and differ only in whether representation plays a role in their means. A deeper review of Ariew's fourfold account of the differences (purposive vs. non-purposive, rational vs. non-rational, intentional vs. non-intentional, and external vs. immanent) will help to clarify what I mean. As I see it, the four distinctions all come down to one thing: whether or not thought plays a role in *the means by which a goal is achieved*.

First, Ariew deems human teleology *purposive*, and that of other natural organisms *non-purposive*. There are two ways to interpret the word "purposive" in this claim. If we take it to mean something along the lines of "for a purpose" or "serving a purpose", then Ariew means that the behaviors of non-human organisms are not for the sake of anything. I find that difficult to agree with and Ariew doesn't defend that claim, but, to me, both the human and non-human-organism cases appear equally purposive in this sense, a point of view I'll examine more deeply in Chapter VII. We can locate a more defensible distinction if we take Ariew's "purposive" to mean something along the lines of "on purpose" or "intended". In other words, if this interpretation is right, Ariew

is telling us that human teleological items and events, unlike those of other organisms, are preconceived in the mind.

Second, Ariew calls human teleology *rational*, and that of other natural organisms *non-rational*. But this is much the same distinction as the previous one. What he is saying is that reasoning and planning go into human goals, and not into those of other organisms. The reason behind human behaviors and artifacts has been thought out, while the reason behind those of other organisms has not.

Third, Ariew calls human teleology *intentional,* and that of other natural organisms *non-intentional*. The common interpretation of the word "intentional" differs from the philosopher's interpretation; however, on either interpretation we end up with roughly the same distinction here. Under the common interpretation, Ariew's "intentional" again means quite the same thing: intended, or preconceived. And if he means it in the philosopher's sense, then he is saying that human teleology is rooted in some mental representation of the world or of the goal. Still, though, we are focused only on the fact that human goal-directedness is preconceived in the mind.

Lastly, Ariew labels the teleology of human artifacts and behaviors as being mediated by external evaluation while that of other organisms is immanent, or internally derived. But what exactly does this mean? The "external evaluation" that Ariew says is required for human teleology means that the judgment of whether or not a behavior or an artifact serves a purpose or fulfills a goal occurs in the human mind, rather than in the behavior or the artifact itself. A knot has a purpose just when it helps to affix some item the way a person wants it to—an evaluation made in that person's mind and thus "external" to the knot. On the other hand, the parts of organisms are purposive by their nature and have no need of an external judge to ensure that they are. As we saw with the previous three distinctions, the human side of this one is set apart by the role of the mind. We are looking at four ways to describe the same characteristic: The teleological aspects of human

behaviors and artifacts comes from a *representation* of some sort in the human mind, while the teleological aspects of non-human organisms' behaviors, products, and processes do not.

Paring down Ariew's distinction like this leaves me with a different view. As I see it, the distinction between what he calls "agent-specific teleology" and "teleology pertaining to natural organisms" is a difference primarily in terms of what means each group employs to achieve their goals—the human version using representation, among other tools, as those means. But both human teleology and the teleology of other organisms seem to me to be *goal-directed* in the same way: there is no fundamental difference here in terms of whether or not the behaviors of humans or other organisms are done "in order to" achieve something or "for the sake of" something. The only difference is whether or not the goals have been represented in a human mind along the way. End-directed items and activities that have not been planned by a psychological agent are nonetheless end-directed.

Dennett (2014) makes the point vividly with a pair of examples.

> Elizabeth Marshall Thomas imagines that dogs enjoy a wise understanding of their own ways: "For reasons known to dogs but not to us, many dog mothers won't mate with their sons." (1993, 76). Nonsense. There is no more reason to think dogs know the reason than that we know the reason why we yawn. There probably is a reason, but we don't know it yet, and it doesn't stop us from yawning. Probably she means something much milder and apparently defensible: she means that we don't know what the discriminated feature is that triggers dog mothers' reluctance to mate with their sons. Well, but we can find out by doing experiments. The first and simplest is to isolate a male puppy from its mother as soon as it is feasible, raise it elsewhere, and return it and see what happens. Will she recognize it? If so, the

> discriminated feature is very probably an odor. *There is a reason* why that odor provokes that aversion, but dogs don't know that reason. (Dennett 2014:56)

> When a low-nesting bird leads the predator away from her nestlings by doing a *distraction display*, she is making a convincing sham of a broken wing, creating the tempting illusion of an easy supper for the observing predator, but she need not understand this clever rationale. (ibid:57)

Dennett's argument is that reasons (and goals) exist whether or not there is someone (or something) that represents those reasons (or goals) (see also Aristotle *Physics* II.8 for an early version of this argument). We may or may not know what purpose something has been done for, or created for, or is used for, but that doesn't mean the thing is not for that purpose. And, furthermore, we usually can—through scientific inquiry—eventually come to understand the purpose.

The female dog has no interest in mating with her son most probably because that would increase the likelihood of genetic and developmental malfunction in her offspring, in the same way that inbreeding causes such problems in human lineages, but the dog doesn't know—and doesn't need to know—that this is why. She simply has no interest in doing it.

As I argued earlier, although humans are largely aware of their goals, there is an important level at which humans, just like dogs and low-nesting birds, are not at all aware of the goals that motivate them. If you ask me why I went to the grocer's, I will tell you I went because I wanted to buy eggs. And if you inquire further to what end I desired eggs, I will say it is because I intend to bake brownies. And if you inquire further still as to what end I have in mind in baking those brownies, the answer will be "to fulfill my hunger and my sweet tooth and my chocoholic tendency, and maybe to please a friend as well". Eventually though, no matter which of my behaviors you

inquire about, my responses will bottom out in emotional descriptions of one sort or another. And if you inquire further to what end I have those emotional incitements that cause me to represent a goal and then to do work in order to follow through on that goal, I will not be able to answer you except in terms of theory. I will no longer be able to say what *I* was thinking—what I was representing in order to achieve my goal—and instead I will have to resort to a description that matches our descriptions of nonhuman teleology. In the worst case, I will simply have to say, much as we surmise the answer from the dog or the low-nesting bird would be, "I don't know; I just want it[21]." Ultimately, human goals are just like other organisms' non-psychological goals; they are "for the sake of self-preservation or preservation of the species" or "for the sake of the organism", as Ariew says of "teleology pertaining to natural organisms" (2002)[22]. Intentional teleology *is* teleology pertaining to natural organisms, and the reason is straightforward: intentional agents, such as humans, *are* natural organisms[23].

Standards

Goals, whatever they turn out to be, seem to come part and parcel with standards of achievement, or "norms of performance", as they are sometimes called. If the goal is to win the race, one must make it across the finish line before other competitors do. If the goal is to get home before the rain sets in, one needs to be dry upon arrival in the foyer. The distinction of whether one

[21] In re-reading Douglas Hofstadter's (2007) *I Am a Strange Loop*, ten years after having first read it, I was surprised to find that the idea expressed in the above paragraph inadvertently, but quite closely, mirrors page 96 of that work.

[22] Well, it is actually a fair bit more complicated than this first approximation admits. Trouble cases that we can eventually analyze include some seemingly counterproductive but occasionally real human goals, such as altruism, celibacy and suicide. An explanation of these cases will have to wait until after we have a theory.

[23] This conclusion might give some readers pause. One concern might center on what we should then make of artificial intelligences. Might they fail to be intentional simply because they are not natural organisms? I don't think that inference follows. I borrowed the term "natural organism" from Ariew's text, but a more fitting term would be "living beings". What it means to be a living being, to be alive, is still up in the air and will be addressed later in the dissertation. But my answer to the question regarding artificial intelligences is: To the extent that machines can be alive (and I think they can) they will also be able to have intentions . . .

is then dry or not is the crucial element that determines whether one has achieved one's goal. At least part of what it means to "have a goal" is to have a "stick" by which to measure a performance.

Philosophers use the term "normative" to refer to circumstances in which norms or standards somehow apply. The difference is often put in terms of the "is/ought distinction" describing the way some judgments in the world are about how things *actually are,* while normative judgments are about how things *ought to be* (Hume 1739). Normative claims compare an object or an event with some type of expectation, and since the measuring sticks that accompany goals play the role of those expectations, goal-directedness is inherently normative. One of the central quandaries of teleology lies in finding a natural, objective basis for this normativity. Where do norms come from? What justifies our saying that something in the world *ought* to measure up in some way or other?

The theory presented later relies on a distinction between two kinds of norms that are worth describing now. The first kind is what can be called *comparative* norms, and while these norms appear to be objective, they are not necessarily associated with goals. For instance, the dwarf planet Ceres ought to continue along its orbit around the sun tomorrow according to the norm set by its having taken that path every day for thousands, if not millions of years now. We simply compare Ceres' performance tomorrow to its previous performances, but whether or not it continues is neither good nor bad; it simply does or it doesn't. Similarly, if all the river rocks that make their way a certain distance downstream from a mountaintop are less than seven inches in diameter, then we can compare a new rock arriving in that location to that normative distribution. The new rock ought to fit in the category, although, once again, whether or not it does is not an evaluative matter. Comparisons of this sort are norms because the class of comparable items serves as the measuring stick, but since that measuring stick is mere similarity, comparative norms have no subjective consequences. The reason we say they *ought* to do something or other—the reason we *expect* them to

perform in a particular way—is only because the similarity in their prior circumstances leads us to predict, by induction, a similarity in their outcomes.

The other major class of norm is what can be called *consequential* or *evaluative* norms. I'll use the latter term. Evaluative norms are not simply measured with respect to previous performances. We might say that "you ought to stay out of the sun to avoid getting a sunburn" or "we ought to use a bucket of ice to keep the fish we catch cool". But the point is not only that you ought to stay out of the sun because staying out of the sun has avoided sunburns in the past; that (merely comparative) claim doesn't capture the spirit of the "ought" in this circumstance. You ought to stay out of the sun because sunburns are *bad* for you—because you don't want to get sunburned. The claim is evaluative. Similarly, it might be true that previously used buckets of ice have kept previous fish cool, but that comparative norm would make more sense out of the claim "a bucket of ice ought to keep the fish cool". The claim that "we ought to use a bucket of ice to keep the fish cool" depends on the evaluative norm that keeping the fish cool is *good*. The stick we are measuring by is whether or not the fish will stay fresh (and thus edible) until we cook them.

Both evaluative and comparative judgments are normative in that they require a judgment against some kind of measuring stick, but for evaluative norms, the measuring stick is defined with respect to some kind of subjective evaluation (see also Millikan 2001[24]). We saw earlier that goal-directedness is inherently subjective; I will argue later that this fact can be explained in terms of a proper theory of evaluative norms. I am not the first to suggest this (see *e.g.* Bedau 1992; McLaughlin 2001; van Parijs 1981); however, of the previous theorists who base their work somehow in value, only McLaughlin (2001) offers a theory of what value is and, although I think he is partly on the right path, I don't think the theory he arrives at has gone quite far enough.

[24] Millikan makes roughly the same distinction as I am here, but she takes the opposite view that teleology is rooted in comparative, not evaluative norms.

Perseverance and Plasticity

What gives us the impression that something is goal-directed? When we watch a cat trying to catch a mouse, or a fly bouncing against a window, or a person throwing a basketball toward a hoop, what is it about their efforts that compels us straightaway to characterize the behavior as striving? In other words, if we are to give a theory of goal-directedness, what observations will the theory cover? What objects or activities in the world will fall under its purview? A culturally intersubjective category—one that has a commonly understood term such as "goal-directed" to label it—will automatically come with some intuitions about what does or does not fit in the category. We all know that the cat, the fly, and the basketball player are goal-directed, but what observable phenomena are our minds latching onto when we make such judgments?

One suggestion that has been made repeatedly is that goal-directed items demonstrate both *perseverance* and *plasticity* (Braithewaite 1953; Nagel 1977a; Russell 1945; Sommerhoff 1969; Wright 1968)[25]. (Most authors use the term "persistence" instead of "perseverance", but I must switch vocabulary here to avoid causing confusion since the word "persistence", used in a slightly different sense, will play an important theoretical role in the theory I describe later.) Nagel puts it as follows:

> One feature is the plasticity of such processes—that is, the goal of such processes can generally be reached by the system following alternative paths or starting from different initial positions. The second feature is the [perseverance] of such processes—that is, the system is maintained in its goal-directed behavior as a result of changes occurring in the system that compensate for any disturbances taking place

[25] See also Bedau (1998) who presents a theory of life as "supple adaptation", which he says is teleological, "not to be equated with natural selection", and which he describes in terms that have a fair bit in common with the notion of plasticity, albeit at the level of a population rather than at the level of an individual.

> (provided these are not too great) either within or external to the system, disturbances which, were there no compensating changes elsewhere, would prevent the realization of the goal. These features can be regarded as identifying marks for ascertaining whether a process does indeed have a goal, and if so what it is. (Nagel 1977a:272)

I cite Nagel in particular because he concludes his description of these features carefully by calling them "identifying marks"[26]. Recognizing perseverance and plasticity in the behavior of an object or organism might help us identify goal-directedness, but such identification is not foolproof, and Nagel recognizes this[27]. The pendulum and the marble in the bowl, for instance, each carry both marks—they are plastic, in that they can be started from a large number of initial positions and still run to the same "end state", and they are perseverant in that moderate perturbations will not sway them from their ultimate course—but nonetheless neither system is typically seen as being goal-directed (see also Bedau 1992a).

Something else is going on in our minds when we intuitively exclude the pendulum and the marble in the bowl from the category of goal-directedness. The division between perseverant and plastic behaviors that are goal-directed and those that are not appears to be similar to the division between "ends" that are the result of goal-directedness and those that are not. There is a qualitative difference between these categories. Bedau suggests the distinguishing factor is *value*—whether or not something potentially benefits from achievement of the goal—and I think he is spot on, but, as I said in the previous subsection, a deeper analysis of that concept will have to wait until later chapters. Whether or not we can find a more consistently applicable theory of goal-directedness, it

[26] A term by which Nagel means "characteristics that are present commonly enough to be used for identification, even if they are not defining characteristics."
[27] Wright, on the other hand, takes these marks more seriously, saying, "What is *essential* to the teleology of a system is the plasticity of its behavior and its persistence towards a goal" (Wright 1968:222, emphasis original).

is worth wondering why perseverance and plasticity characterize it at all. That is to say, why are these qualities commonly and intuitively associated with goal-directedness? I will try to address this question near the end of the dissertation.

An Ambiguity in the Interpretation of "Teleology"

When we speak of, or theorize about, purposes or teleology or goal-directedness, there are two main themes that we might be speaking of that overlap under certain interpretations but that can be seen under other interpretations as entirely distinct. Confusion between these themes may muddle up both modern conversations and reviews of historical debates, and so it will be useful to recognize the potential distinction between these themes before setting out.

The first meaning of "teleology" refers, at a grand scale, to the purposes for which the arrangements of the universe have been made. There is a potentially teleological question that can be asked about why the things that we observe in our world are the ones that exist rather than other imaginable ones, and about how the universe has come to be ordered in the ways that it is (which might seem to be good or purposeful or designed). Proposed answers for such a question are usually given in terms of some unobserved metaphysical and universal teleological force—often, but not always, a deistic force. Some writers have called ideas of this kind by the name *cosmic* teleology (see, *e.g.*, Mayr 1974/1988, 1992). As we'll see in chapter III, the history of cosmically teleological thought begins with Plato's conception of ends and to this day remains a mainstay of creationist thought.

The second meaning of "teleology" regards the purposes or goal-directedness of individual items in the world. When we observe certain behaviors of items in the world, and the existence of certain other objects, we get the intuition that these things are behaving as if they themselves are

goal-directed or that they exist primarily thanks to their relationship with some other observable goal-directed agent. This concept is often termed *immanent* teleology, which simply means teleology that somehow arises from and inheres in an object itself. We'll see in chapter III that this more local-scale concept of purposiveness begins with Aristotle's impressions of ends. I wish I could say that immanent teleology is the popular view today, but, in fact, it is not. The most common modern scientific view, as we'll see in chapter VII, is the homocentric, anti-teleological view, neither cosmic nor immanent, in which most items that appear to be behaving in a goal-directed manner only *appear* to be behaving in a goal-directed manner, while the only *real* immanent teleology in the world arises from the intentions of the human mind.

Now, having described these two interpretations of the word, I can clarify my own project: I will be attempting to describe an immanent teleology in terms of the structures and relationships found in and between organisms and other biologically-relevant patterns—structures and relationships that can account for their goal-directedness and functional natures. I believe it most prudent to try to explain these phenomena in terms similar to those with which we try to explain everything else in our universe: simple, physical, and mathematical terms based only on what we are able to observe. I will set this view in opposition to both the homocentric view, which I consider to be a fallback position taken by those who don't see an obvious way to explain immanent teleology, and the cosmic view, which I consider to be both false and explanatorily incomplete due to metaphysical assumptions. I'll say a few words about the latter point now.

I don't find it justifiable to believe that there is a creator that designs organisms or that organizes or intervenes in the universe in other particular ways. Nor do I think it reasonable to believe that the universe itself has a grand purpose (whose purpose would it be?), nor that evolution is directively organized toward some particular end, nor that there is any other kind of cosmic-level teleology that needs to be explained. Cosmic teleology, for me, is an unviable proposition because

the purposeful behavior we really observe occurs only—*only*—at the levels of organisms, their artifacts, and their societies. With the sum of the achievements of science up through the twentieth century as a backdrop, there simply are no observed patterns remaining at any other scale still calling out to be explained in purposive terms.

That said, there is one possibility that I find vanishingly unlikely but that I also can't entirely rule out and so I must mention it briefly: The possibility that while the details of our world are not managed by a creator, the entirety of our world is a grand artifact created by a grand artificer (*i.e.* a god) who has simply set up and seeded the fundamental physical constituents and constants that comprise it. Our universe might, for instance, be a computer simulation of some sort—an extravagant digital dollhouse residing in a machine that occupies only a tiny corner of a much larger universe. If this were the case, my project would still hope to explain only immanent teleology. The conundrum of how teleology arises in our world, given the physical laws (of the simulation), would remain unchanged. We would merely have to answer the *additional* question as to how the creator came to be a teleological agent in its own world.

B. A Note on Methods

Until now the traditional method of approach . . . has been a highly intuitive one. A set of (intuitively) acceptable examples . . . are offered and an analysis is constructed to meet those examples. Counterexamples are then sought and found . . . Throughout the entire procedure, the putative counterexamples themselves must be accepted or rejected on purely intuitive grounds, with no clear agreement as to whether they should count as genuine cases . . . or not. This trial and error method may indeed be the best we can do.

—Fred Adams (1979), on "function".

The test of examples and counterexamples is important. Yet in this case . . . there is a risk that it will decay into the dull thud of conflicting intuitions.

—John Bigelow and Robert Pargetter (1987)

The work in this dissertation is a theoretical approach to a set of interrelated topics that, traditionally, have been addressed chiefly by philosophy. The attempt I am making to edge those topics from the domain of philosophy toward the realm of science can be divided approximately into two parts.

Subjects and Theories

In much of the first part (predominantly, chapters I through VII), I will use some traditional philosophical tools to outline my subject. One of the earliest challenges when offering a theory of a

phenomenon is to find a way of drawing the figure that is one's topic out from the ground that surrounds it. One of Newton's great accomplishments when developing his theory of gravitation was his realization—not nearly as simple as it sounds to us now—that the moon's orbiting the earth, the planets' taking their paths around the sun, and apples falling from the branch to the ground are all species of motion that can be attributed to the same cause, while other types of motions need not be accounted for under a single explanation. Analytically unifying the particular phenomena that he did allowed Newton to see a more clearly outlined subject for which he could then develop a single explanatory hypothesis—the theory of universal gravitation. Without that unifying insight, it would have been far more challenging to understand each of those kinetic phenomena separately. A primary role of philosophy, as I see it, is to try to find ways such as this of divvying up the myriad overlapping patterns that we find in the world and usefully grouping them to suggest cohesive topics for science to investigate.

The division of labor I am describing here may seem to have much in common with Carnap's (*e.g.*, 1936) view that philosophy is tasked with the job of defining scientific concepts before science can investigate them, but I want to distance myself somewhat from that comparison. Carnap saw philosophy as providing *a priori* definitions. What I am advocating, instead, is a philosophy that provides analyses—not definitions but *rough characterizations*—that give theoretical science a baseline from which to work, but that may also be revised by the results of both theoretical and experimental science. There are no deductively *a priori* claims in the method I am advocating and the relationship between philosophy and science is one of interplay, of back-and-forth, with mutually constraining epistemic dynamics.

I will approach the task of outlining my subject in Part I by using, among other things, a reserved form of the philosophical method of conceptual analysis that I will call *cautious* conceptual analysis in order to contrast it with another sense of the term that we can call *conventional* conceptual

analysis[28]. I make this distinction firstly because I have found that, depending upon which philosopher you ask, you may hear two quite incompatible ways of interpreting the details of what is meant by the term "conceptual analysis", and secondly, because I wish to highlight important differences between, on the one hand, the now oft-criticized conventional version of analysis and, on the other hand, my own work, which, without the distinction I am making, might be found to resemble that conventional version due to the use of some similar tools of reasoning.

Conceptual analysis, in its general form, is the act of trying to reason about the world by using the structure of our concepts as a guide. Philosophers reflect upon their own concepts, their usages of them, and their intuitions about what they refer to, in order to find out what may be revealed by those reflections. The difference between the cautious and conventional forms lies in what the practitioner believes they are justified in doing with those reflections. The conventional interpretation is usually understood as being capable of *deducing a theory* while, as described above, the cautious interpretation sees the method as merely *producing a subject* . . . to which theory, and eventually experimentation, can later be applied (see also Jackson 1998).

The second part of the dissertation (chapters VIII through XIV) will more closely resemble theoretical science in that it will attempt to offer a minimally ambiguous picture of an underlying structure that may account for the subject outlined in Part I. Because of this, Part II is less methodologically controversial (though it may still be theoretically controversial). When a theory has been presented, it becomes a target for attempts at both falsification and predictive confirmation, two processes that are widely accepted as capable of adjudicating scientific matters, even if only provisionally so. What I offer in Part II may be right or it may be wrong, but the process of determining which of these is the case is, in principle, approachable.

[28] Jackson, 1998, makes a very similar distinction, naming the two styles "modest" and "immodest" conceptual analysis.

So, leaving aside part II for now, I want to say a little more about the methods used in part I. In this section, I don't intend to give an extensive defense of conceptual analysis, but I will very briefly describe what I am calling the cautious version of it, as well as advocate for a reduction of impulsive hostilities against the tools of analysis when those tools don't necessarily reflect any commitment to the goals and entitlements presumed by the conventional version.

Conventional Conceptual Analysis

The method of conceptual analysis is as old as philosophy itself, though the debate over its validity or usefulness seems to stretch back only a handful of decades. On one hand, there are those who find the method to be an utterly indefensible technique and perhaps the largest historical blunder in the field of philosophy (*e.g.* Colin Allen, pers. comm.; Laurence and Margolis 2003; Millikan 1991; Stich 1993). On the other hand, there are those who find the method to be an utterly indispensable technique—the absolute lifeblood of philosophical inquiry (*e.g.* Bealer 1987, 1998; Chalmers 1996; Dennett 2007, 2013; Jackson 1998; Lewis 1994). The disagreement is fiercely polarized, but I believe that the two groups are talking past one another by using the same term to convey different visions.

The opponents of conceptual analysis deride it as being an essentialist search for necessary and sufficient conditions, in a world that is usually not composed of essentialist patterns (I'll describe and explore essentialism in a few moments). They also scorn it for being performed by using intuitions about our hypothetical categorization of hypothetical examples—that is to say, the practitioners of analysis often imagine a situation, and then ask themselves whether they would or would not intuitively consider that situation to be an example of the category or concept they are analyzing. The opponents of the method point out that the fallibility of our intuitions and of our

categorization abilities are both obvious and well documented, and so it would be folly to depend upon them in deducing theories (Kripke 1972; Putnam 1962). These are serious charges that certainly justify skepticism if indeed the goal of conceptual analysis is one of deducing an essentialist theory. But I will argue that modern proponents and practitioners of conceptual analysis can (and do) still analyze concepts without any necessary obligation to the essentialist onus, and, I suggest, if they do so cautiously—if they remain aware of the fallibility of our intuitions and categorization abilities—they can tread lightly enough to avoid drawing premature conclusions.

Essentialism

The conventional form of conceptual analysis was first employed, so far as we know, by Socrates. A Socratic definition is meant as a kind of a theory; a definition is given in terms of what are now called the necessary and sufficient conditions for a phenomenon. The ideas of necessity and sufficiency are invoked as limits such that just those conditions that make up the definition, and no more (thus, all together they are sufficient) and no less (thus, they are each individually necessary), are required for an item to fit the category of theoretical interest.

For instance, in Plato's *Theaetetus*, Socrates discusses Theaetetus' proposal for a theory of *knowledge* as "true belief with an account" (*Theaetetus,* 201-210), or what is now sometimes called "justified true belief" (JTB).[29] On this view, knowledge consists of 1) beliefs that 2) are true, and that 3) are justifiably believed to be true (see also Ayer 1956; Chisholm 1957; Gettier 1963; Plantinga

[29] For other places where Socrates searches for essential definitions, see, *e.g.*, Plato's *Euthyphro* (piety), *Charmides* (temperance), *Republic* (justice), *Meno* (virtue), and *Hippias Major* (beauty). Knowledge as justified true belief is also discussed in the *Meno* 98a2; *Phaedo* 76b5–6, 97d-99d2; *Symposium* 202a5-9; *Republic* 534b3-7; and *Timaeus* 51e5

1993). Since these conditions are meant to be necessary and sufficient, nothing else is relevant; we have knowledge as long as we have justified true beliefs.[30]

The pursuit of necessary and sufficient conditions for a phenomenon can also be cast as the search for the *essence* of that phenomenon and therefore analyses that purport to provide such conditions are termed "essentialist".

The Trouble with Essences

As it turns out, however, not everything is amenable to being described in essential terms and, in fact, it appears that very few things are. We ran into the difficulty, already, when trying to define "ends" in the previous section. There we found that what counts as being an end is not an objective matter for which we could give unexceptionable conditions, but instead it is somehow relative to subjective interests.

Wittgenstein, in his *Philosophical Investigations*, first drew attention to the difficulty of providing essences for our concepts. His most famous example was the notion of a *game*. In the broad array of types of things that we think of as games there doesn't appear to be an essential set of necessary and sufficient criteria, much less even a single criterion that is common to all of them. Instead, Wittgenstein says, there appears to be a "family resemblance" relationship in which one game may resemble some others in some regards, while those others resemble still others in differing regards, creating a chain or network of relationships, with no single thread running through every instance (Wittgenstein 1953; see also Fodor 1981; Hofstadter and Sander 2013; Lakoff 1987; Rosch 1973; Vygotsky 1986).

[30] Socrates ultimately argued that the JTB theory isn't quite right, and many contemporary theorists have done so as well, but no one in analytical philosophy has yet given a better answer.

If one looks around one finds quite quickly that the same difficulty in essential categorization applies to concepts as diverse as "*band*, *chair*, *teapot*, *mess* and *letter 'A'*" (Hofstadter and Sander 2013:pp) as well as *species*, *rabbit*, *Australopithecus*, *Homo*, *African American*, *dead*, *alive*, *murder*, *poverty*, *Republican*, and *Democrat* (Dawkins 2014).

Things with Essences

Despite the troubles with essentialism, there are a few ordinary patterns that do appear to be describable in terms of rigid essences. Triangles and circles as ideal mathematical forms are commonly cited examples (*e.g.* Mayr 2002). It is not at all blurry that a triangle has just three straight intersecting sides or that it is made of lines that meet in just three angles in a Euclidean plane. Either of those definitions seems to provide necessary and sufficient conditions. In addition to geometric forms and some other mathematical concepts (such as *pi* or *prime number*), we find that the subjects of at least a few scientific concepts appear to have essences, for instance: *electron*, *proton*, *tungsten,* and even possibly *DNA*. I'll address the grounds on which this may be debatable in chapter II, but it seems we can say with certainty that if a thing has 74 protons bound together in a single nucleus, it is tungsten, and that if a material made up of a particular isotope of tungsten (having a certain number of neutrons and another specified number of electrons) then it will behave in certain very regular, predictable ways in terms of such things as its electromagnetic properties, thermal conductivity, melting point, density, and so on. There is something very much like an essence to fundamental particles and various specific isotopes of elemental substances. Philosophers typically call these types of things by the term "natural kinds", because they seem to be categories whose

identities are created by nature, rather than by our minds—they seem to be not just subjectively similar but also objectively similar and thus not subject to Wittgenstein's concern[31].

So what is the difference between the majority of concepts that are blurry and those few for which we can find a rigid essence? Under what umbrella could we possibly unite *DNA*, *tungsten*, *protons* and *triangles*, while excluding *rabbit*, *chair*, and the *letter 'A'*? This is the essentialist question turned in upon itself. Essentially, it is asking, "What is the essence of having an essence?" Later, I will suggest some lines along which we may be able to divide these categories.

Cautious Conceptual Analysis

If we relinquish essentialism, or assume at the least that the particular topics we are interested in are not necessarily natural kinds, there are still two roles that I think the tools of conceptual analysis can play in philosophy. The first, to repeat what I said above, is that reasoning about our concepts and thinking about example cases can be a useful tool not so much for deducing a theory, but for helping to produce a subject about which one can then suggest—and eventually test—a theory.

The subjects of science are not handed to us pre-packaged. When we are attempting to characterize a phenomenon, we need to come up with reasoned categories about which to theorize, based on various observed similarities, just as Newton had to do when he analytically selected certain types of motion while deciding to exclude other types. A primary way of discovering and recognizing those similarities is to reflect upon our own concepts and the observed examples from

[31] However, a good number of philosophers also count as natural kinds such categories as *tree* and *rabbit*, which they think are also handed to us by nature. I am inclined to agree with Hull (1965) and Dawkins (2014) in regarding such biological concepts to be subjectively labeled human categories, far more like *chair* than *proton* (but for a contrary view see Griffiths 1999; or Okasha 2002). It is clear that we should want to categorize all rabbits as a kind—*rabbit* is a useful category for both biological and practical use—but it is not at all clear that there is a *natural* kind. There just don't seem to be any necessary and sufficient conditions that could cleanly capture the set of all rabbits without getting blood and fur all over the place.

which they are derived. As Frank Jackson (1998) puts it: the interesting philosophical questions are not ones that have unfamiliar definitions, but ones that inquire about subjects "*according to our ordinary conception*". When we ask, for instance, whether a subject such as Jackson's example of free action really exists, what we are interested in finding out is whether the things we think of when we hear the term "free action" exist. The very question is about a pattern we take ourselves to perceive from the world. And, in order to identify our ordinary conception, Jackson points out, the only possible way is to "appeal to what seems to us most obvious and central about [for example] free action, . . ."—that is, to use our intuitions. We have no other basis from which to begin. Of course this doesn't mean our *scientific theory* of the phenomenon needs to be derived from our intuitions—indeed it should not—it only means that our common theory (philosophers say "folk theory") of the phenomenon is derived from our intuitions and that our scientific theory, in one way or another, ought to account for our perception of that folk theory.

Our minds reflect the world imperfectly, much like the reflections in a funhouse mirror, but a careful philosopher who is aware of this risk, and who is armed with some idea of the shape of that mirror (as produced by the nascent yet fast-growing field of cognitive science) can proceed to characterize what they see reflected in their concepts in order to get a rough idea of the subject they plan to theorize about, and in order to cast aside patterns that they think more likely to reveal the shape of the mirror than the shape of the world (see also Bacon 1620; Hofstadter 2007).

Counterexamples

The second role I see for the tools of conceptual analysis, used cautiously, is the common activity of employing counterexamples in order to discover insufficiencies in the mapping or translation between a theory and a subject. When a theory of a phenomenon has been offered,

through whatever means, a clear counterexample can at any time suggest either that the theory needs revision in one way or another, or that the subject that the theory is meant to account for was not consistently outlined.[32] Counterexamples (as long as they are not too far-fetched) are observational data. That they were not necessarily *actually* observed in the present does not invalidate them if they are clear, central, obvious cases from memory or straightforward extrapolation (see also Hurley, Dennett, and Adams 2013).

Using counterexamples to question the correspondence between a theory and a subject does not have to be an essentialist activity. Modern cognitive science has given us a number of alternative notions of blurry-edged rather than essentialist concepts that are nearer to Wittgenstein's portrayal of family resemblances. For instance, one might think of concepts as being composed of prototypes and their variants (Rosch 1973) or a series of exemplars (Smith and Medin 1981), or as having either blurry conceptual halos and a dynamic underlying structure (Hofstadter and FARG 1995; Hofstadter and Sander 2013), or non-necessary bundles of common features (Prinz and Clark 2004)[33], or necessary along with exceptionable conditions (Jackendoff 1983). If we characterize our subject in any of these ways, then a counterexample can still highlight a way in which a theory that is meant to account for that characterization may fail to (although determining whether that failure is a fault with the theory or a fault with the characterization may require further work).

[32] The way I see it, the topics produced by a cautious conceptual analysis are equally as subject to revision as theories are, and the process of hunting for a theory is one of interplay and mutual revision, but with a focus on broadening the subject as much as possible, while narrowing the theory as much as possible in order to strive to find as much generality in our theories as we can.

[33] See also Boyd (1999) who also used a notion of non-necessary bundles of features (he calls them *homeostatic property clusters*) to try to redefine species concepts, such as *human*, *lily,* or *tobacco hornworm*.

Summary on Methods

The critic of conceptual analysis who finds my use of these methods disconcerting should take note that, by the end of Part I, when I have finished mulling over many cases and counterexamples to theories, I will not yet have offered anything that looks like a theory. I will not have presented a list of necessary and sufficient conditions for either *goal-directedness* or the related concept of *function*. Instead, what I will present (in Chapter VI) is simply a list of questions and phenomena related to these concepts, and some reasoned commitments as to what patterns I think any theory of these subjects should be attempting to account for. I make some strong, and probably controversial claims; but they should be controversial for what they say, not for the methods by which I have developed them.

C. Organisms and Artifacts

The origins of AI, in one form or another, go back to the perennial pursuit of human beings to understand their place in the overall scheme of things, both as creations *and as* creators.

—Hamid Ekbia (2008, emphasis added)

As by now you will have recognized, the domain of teleology appears limited, more or less, to organisms and their artifacts. This state of affairs has been a source of both insight and confusion in teleological inquiry, as I will explore in detail in later chapters. But I raise the point now for a different reason: I began my own foray into this field because of my interest in artificial intelligence (AI). In imagining what it would take for modern or future scientists to build a real, general, human-level AI, I was inspired by Hamid Ekbia's (2008) *Artificial Dreams*, which explores some of the reasons AI researchers' achievements, to date, have always fallen far short of their most ambitious dreams.

Ekbia's book highlights a number of fundamental tensions in those dreams, one of which really caught my attention, and which I think is best put in terms of the contrast between two categories that have very different ways of being purposeful: organisms and artifacts (see also Collins and Kusch, 1998, from whom Ekbia draws some of his discussion).

On the one hand, AI researchers are attempting to build something like an organism—something willful, lively, and internally motivated; a creature that behaves on its own behalf; an autonomous system that seems to have, and to serve, its own purposes. On the other hand, each of these researchers is building, by definition, an artifact—a human-made device that we understand to serve the purposes of another autonomous agent (typically, the researchers themselves). As Ada

Lovelace put it in her own critique of artificial dreams, over a century before the term "artificial intelligence" was coined, "The Analytical Engine has no pretensions whatever to originate anything. It can [only] do whatever we know how to order it to perform" (Menabrea and Lovelace 1843)[34].

It takes some effort to understand Lovelace's dictum clearly. Certainly, today, we know that machines can in fact compose music and draw pictures that their designers had never heard, hummed, seen, or imagined (see *e.g.*, Cope 1996; Ekbia 2008; McCorduck 1991) and so, in some sense, these machines are "originating" something, in spite of Lovelace's claim.

However, the partially-novel products fashioned by these machines are created within bounds that are specified by the machines' designers and, more importantly, the so-called creative acts are performed only in order to satisfy *the goals* of those designers. The purpose for which Harold Cohen's drawing machine Aaron creates its drawings is that Cohen (or another user) wants to produce a machine-drawn illustration and, when Aaron is started up, it dutifully draws a picture by following a complex set of rules that Cohen has injected into it, with a number of randomized parameters. If Cohen never puts spaceships into Aaron's repertoire, Aaron will never draw a spaceship. More importantly, Aaron will never *want* to draw a spaceship. In fact, Aaron doesn't ever want to draw anything at all, because wants are not a part of Aaron's architecture. The desires and intentions behind Aaron's works are Cohen's, as are the style, the repertoire, and even the semi-random artistic choices. As Ekbia notes, any apparent autonomy found in an artifact comes to an end when you dig deep enough (again, see also Collins and Kusch 1998). At some point, there is another agent behind the scenes ultimately in control of the machine in one way or another, telling it which destination to drive to, describing what types of elements may show up in its drawings, changing its parts when they wear out, and, most importantly, perhaps, defining the machine in the

[34] The Analytical Engine was a design for a mechanical general-purpose computer described by Charles Babbage in 1837. The machine, were it to have had all its kinks worked out and had it actually been built, would have been what later would come to be known as a Turing-complete—or universal—computation machine much like any of our modern devices, though quite a bit slower due to its purely mechanical nature.

first place as either an autonomous vehicle, a drawing machine, a music composition machine, or a chess, go, or *Jeopardy!* player (see, *e.g.* IBM's Deep Blue and Watson or Google DeepMind's AlphaGo).

How, then, can something that is human-made, something that is clearly an artifact, something that derives its purposiveness from its creators or its users . . . how can it also be an organism? How can it be something that has its own purposiveness? When we eventually build an artificial intelligence, which will it be: organism or artifact?

Some may see this tension as an argument that artificial intelligence is a fool's errand, but I think that conclusion would be premature. I take it, instead, to signify the fact that we simply don't yet know enough about what it means to be an organism or an artifact—we don't yet know enough about how either of these types of things come to be purposive. One of the fundamental questions that need to be clearly—even mathematically—answered before our scientists and engineers can truly work in earnest on building an artificial intelligence is the question of how any particular parcel of particles can come to be purposive.

A Brief Preview

In this work I hope to present a view of purposiveness that not only can differentiate between organism and artifact and clarify how the fields of artificial intelligence and artificial life can usefully progress in attempting to build organisms while at the same time necessarily building artifacts, but that also can underlie a theory of life, agency and autonomy. As I described in the previous section, the view presented comes in two parts, which represent two almost-separate projects. In the first project, I attempt to recast the subject matter of teleology by describing what I

think are the relevant patterns to be accounted for, while setting aside those patterns that I think are irrelevant or illusory. That project begins here and concludes by Chapter VII.

Chapter II will outline some of the physical and metaphysical assumptions that underlie my interpretation of the world as applied to both parts of the dissertation. This is a work in the materialist tradition of science and philosophy, and so my goal is to locate theories of subjectivity, value, identity, and goal-directedness in the physical world by resorting only to physical constituents, mathematical relations, and other phenomena that can emerge from (or, in Jackson's 1998 sense, "are entailed by") the physical constituents of our world.

In chapter III, I review the history of teleological thought in some detail, tracing it from both its psychological and historical roots in animistic beliefs through the annals of philosophy and biology up till the end of vitalism in the early twentieth century. Throughout that time, many insightful issues were raised, and it is worth keeping them in mind so that our later theory can be held accountable to the many observed facets found during early observations and analyses of life and goal-directedness.

In chapters IV and V, respectively, I analyze issues surrounding the concept of *function* and then review most of the modern theories and analyses of function. Function is an important teleological concept that has been the focus of the vast majority of recent teleological inquiry, and, despite my own disbelief in the field's central notion of "proper functions", the contributions made in this recent body of literature have laid many of the foundations for the view of teleology that I will later advocate.

In the very brief chapter VI, I will review the progress made up to that point by relisting the many questions raised by the conceptual analyses in the previous chapters. I take it that these questions form a kind of checklist that both the theory I will advocate and *any* future theory of teleology should be held accountable to explain.

Before giving an account that attempts to answer those questions, I take a bit of time in chapter VII to motivate my work in Part II, by showing, firstly, that the modern intellectual atmosphere has a thorough distaste for goal-centric (instead of function-centric) teleological thinking and, secondly, that this anti-teleological attitude is unfounded, and has, for decades (if not for centuries), hindered scientific progress concerning the foundations of biology and agency.

In Part II of the dissertation, I describe a hypothesis that is meant to account for the patterns outlined in the first part. Chapters VIII, IX, X, and XI are the central description of this account of teleology, including exposition of ideas on identity, value, subjectivity, vitality, and organization. The account is not entirely novel. As you'll see, other theorists have offered most of the pieces of the theory at one point or another. What is novel about my contribution will primarily be the way in which the pieces are arranged together to paint a broader picture.

Chapter VIII will present an overview of how identity and value can produce naturally normative, teleological behavior. Chapter IX will present an account of how a set of patterns in our world can, together, have an identity that may serve as the locus of subjectivity, the beneficiary of value, and the teleological agent that can be said to possess its own goals (and to strive for them through its actions). Chapter X will further develop that account in order to provide a measure of how such an identity could benefit, making certain objects or events valuable to it. Chapter XI will then attempt to use these notions of identity and value to create a system of classification for the various kinds of patterns in our world in terms of their relationships to value.

Various subsets of my readers will be familiar already with varying parts of the analysis, but hopefully the introductions above will help readers to focus on the parts they will find most interesting, while avoiding introductions to topics with which they are already familiar. To understand the core of my thesis, a reader who is willing to accept the reality of goal-directedness

should focus on chapters II and VIII through XI. The reader who doubts the existence or importance of goal-directedness in our world should add Chapter VII to that list [35].

[35] While I'm remarking on what one need or need not read, let me use a footnote to say a few words about footnotes: Although, as you've seen, there are a considerable number of footnotes in this work, the majority of readers can be assured that nothing major will be lost by skipping over all of them. A large proportion of the footnotes are there primarily to forestall certain interpretations, upon which philosophers of various stripes might be prone to base preemptive dismissals of my work. Retaining comments of this type yet moving them into footnotes is my way of attempting to reach a broad audience that contains both philosophers and non-philosophers. While I want this work to be subject to philosophical scrutiny, my primary intended audience is students of cognitive science and AI, like myself just a few years ago, who have their own "artificial dreams" that they'd like to follow.

Chapter II

Physics and Metaphysics

Metaphysics, n.: The branch of philosophy that deals with . . . questions about being, substance, time and space, causation, change, and identity (which are presupposed in the special sciences but do not belong to any one of them) . . .

—Oxford English Dictionary

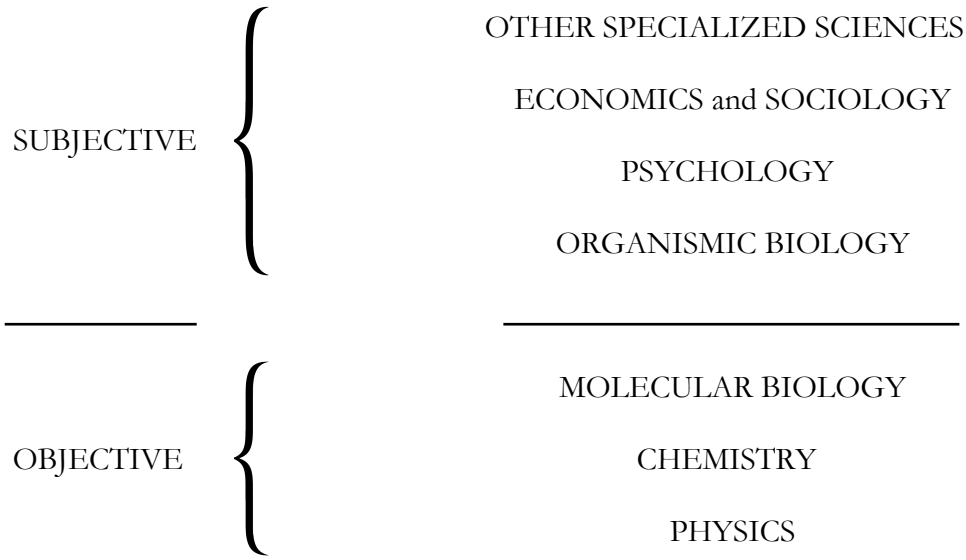
In many people's minds, there is a hierarchy among the sciences, beginning with physics at its base, followed by chemistry, and then proceeding through biology to psychology, anthropology, sociology, economics and so on. In some way, I find this picture compelling. If there were no atoms, there would be no chemical reactions between them. If there were only atoms, but no molecules, there would be no biochemistry. If there were atoms and chemicals, but no organisms, there would be no minds, no cultures, no societies, and no economies. The hierarchy exists in large part because the objects of study at each level within it depend upon the existence of—and the relationships between—some of the objects of study at the levels below.[36]

For the moment, it is not the hierarchy itself that I want us to be concerned with as much as an oft-drawn distinction within it, between what are commonly called the "hard" or "natural" sciences and the "soft" or "special" sciences. These labels have come to be outmoded today, as the previously "soft" methodologies in the so-called soft sciences have largely caught up, in terms of rigor, with those of the hard sciences. Even so, there is still something about the subject matters of

[36] Some see this hierarchy as supporting a reductionist's structure of the sciences, in which each level of the hierarchy consists of phenomena that are merely made of those below. I don't find that viewpoint very compelling; instead I subscribe to the emergent perspective, in which new phenomena often arise at each level not merely from the aggregation of constituents of the levels below, but from the particular organization of those constituents at each level. I will discuss reductionism and emergence further and, in particular, the question of whether biology can reduce to chemistry and physics in Chapter IV.

the various sciences that keeps the hard–soft (or natural–special) distinction alive in our minds. In particular the subjects of the hard sciences relate to *objects* while the subjects of the soft sciences relate to, well . . . *subjects*. Beginning somewhere in the middle of biology and moving upwards through the hierarchy, the subject matter slides from a study of the objective behaviors and interactions of materials in our world to a study of the subjective behaviors and interactions of organisms and agents.

The hierarchy, with this division inserted, looks something like the following.[37]



But there is a problem with this picture. The gap there in the middle, dividing biology in two, represents something important . . . and something *missing*. Really, the divide is between those topics of biology that are smaller than whole organisms (for instance, biochemistry, molecular biology, and genetics) and those topics of biology that involve the behaviors and interactions of whole organisms

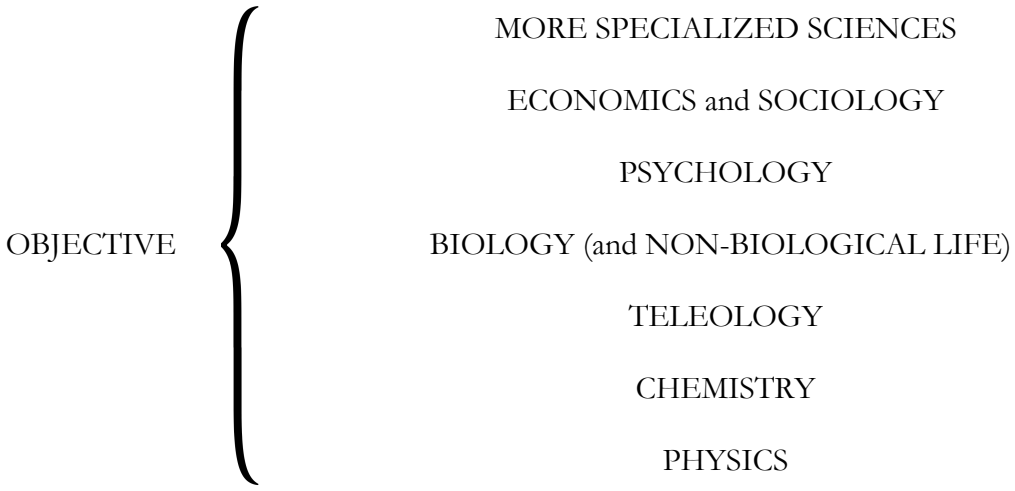
[37] This is an oversimplifying image because, for instance, the boundaries of some special sciences, such as geology and oceanography, cut across these categories and thus have no obvious single location that they might fit in a hierarchy of this sort. However, the picture I am drawing is not meant to capture the categorical structure of all sciences, but rather what is (in popular discourse) a particularly salient cross-section of that structure, relating the biological and humanistic realms to their underlying physical constituents.

(for instance ecology, ethology, and evolutionary biology). And this is an extremely odd thing for the study of living things to do, because it says that, on one hand, we can study the non-living parts of living things and, on the other hand, we can also assume the whole-hog existence of living things, and then study their behaviors and interactions, but there is an important sense in which neither of these activities is actually the study of living things. What is missing is an understanding of how the pieces in the objective branches become alive, so that there is something to study in the subjective branches[38]. The explanatory gap has been explored by a small but important group of theorists whose ideas we will discuss in the second part of the dissertation, but in the mainstream of science, and even in the bulk of biological study, it is largely ignored.

One vision for what could fill this gap is a science of teleological patterns—a science that, firstly, assumes the objects of physics and chemistry along with those of the objective branches of biology; and that then uses the relations between those objects in order to provide an account of vitality, goal-directedness, agency, function, identity, and value, all of which can thereafter be assumed to exist by the more subjective (and subject-based) branches of biology, as well as by other higher-level special sciences.

Now there is clearly much work to be done in order to substantiate the claims of that vision but, assuming that that work can be done, this arrangement suggests a new picture of the hierarchy of the sciences that looks like this:

[38] One might be reminded, here, of "books that told me everything about the wasp, except why", found on the list of "Useful Presents" given in Dylan Thomas' (1952) short story *A Child's Christmas in Wales*.



What I think is beguiling about this alternative arrangement is that, rather than splitting biology between the objective and the subjective, or inappropriately shoehorning all of it into one or the other of these major categories (while still leaving the other so-called soft sciences in the still-unexplained realm of the subjective), the dividing line itself is instead replaced entirely by a study of subjectivity in objective terms, thus bringing everything above the line into the fold of what previously lay below it. The world is objective, even where it is subjective.

I don't expect this vision to convince many readers just yet. As we'll see in Chapters III and VII, many scientists today believe teleology to have been carefully eradicated from modern science. An attempt to reintroduce it (and so very close to the physicochemical foundations of the hierarchy, at that!) may not be well received. But when we recognize that the traditional hierarchy ignores what accounts for subjectivity and that, as was noted above, the biologist's study of life is generally agnostic about life, it becomes clear that something needs to span the divide.

The modern theory of teleology, pieces of which have been coalescing in recent decades, and which I think is described most fully now in the latter part of this dissertation, can begin to fill the gap, because it suggests that it is just at this level in the hierarchy of the sciences that a particular type of arrangement of matter—a particular type of pattern—is able not only to become goal-

directed but also, at the very same time, to come to have an identity along with the prospect of being a potential beneficiary. Moreover, as we'll come to see, I think this particular type of physical pattern is the very thing that logically underpins the two cornerstone processes of biology—cell-maintenance and replication—wherein objective, mechanical, biochemical processes give rise to the kind of survival-and-reproduction machines that give us the impression of vitality, and to which we attribute subjective qualities. If I am right about that, then this physical pattern is an obligatory prerequisite for the existence of the living, agentive, teleological objects under study both in the subjective branches of biology and in many of the special sciences.

My thesis is that it is teleological patterns that form the very nexus between the objective and the subjective. Teleology is the singular thread that weaves together molecules, microbes, minds, and machines into one continuous universal tapestry. And in occupying this unusual position, it crosses paths with a broad array of physical and metaphysical phenomena, coming into intimate contact with, on the one hand, physical phenomena such as energy, entropy, and information, and, on the other hand, many notions of metaphysics such as existence, causation, value, and identity. Because of this, we need to spend some time now introducing the relevant pieces of all these topics.[39]

[39] Much of what follows in this chapter is an inherited view of the world, derived from the modern culture of science and philosophy. In particular, however, I am greatly indebted to two of my academic advisors, Douglas Hofstadter and Daniel Dennett, who have influenced my version of this heritage the most. Although the majority of that influence has come from personal communications, the reader who would like to trace some published pieces of that influence can look first at Hofstadter (2007) and Dennett (1991) and then at the remainder of the works by both of these thinkers, as cited in the bibliography.

A. The World We Live In

The first principles of the universe are atoms and empty space; everything else is merely thought to exist.

—Democritus of Abdera[40]

. . . everything that animals do, atoms do. In other words, there is nothing that living things do that cannot be understood from the point of view that they are made of atoms acting according to the laws of physics.

—Richard P. Feynman (1963)

Much of what I am going to discuss depends on some basic assumptions about the properties and constituents of the physical world and the mechanics of causation therein. The simplest way to state those assumptions is to say that I am working in the framework of what is often called "the modern scientific worldview", a synthesis centered on the canon of eighteenth-, nineteenth-, and twentieth-century physics and chemistry. There is a little room for interpretation about what is meant by this, but I think it is broadly understood to mean what I intend by it.

To be somewhat more specific, however, I can begin by saying this: I will neither help myself to any theoretical substances or forces that haven't been clearly demonstrated to exist by experiments that can be repeated by modern physicists, nor will I tolerate it when other theorists do so, either explicitly or implicitly. While particle physics still faces challenges in explaining the full range of observations that scientists have made (see, *e.g.*, Blum *et al.* 2013; Carlson 2015; Lees *et al.* 1970; Lykken 2010; Persic and Salucci 1992; Pohl *et al.* 2013; Rubin *et al.* 1980; Sushkov *et al.* 2011;

[40] As cited in Diogenes Laertius IX, 44-45 (Hicks 1925).

Trimble 1987), there is a broad swath of theory that is known to describe the world very accurately for the vast majority of modern practical purposes, and it is this theoretical basis that I will assume.

Anchoring that body of knowledge is what is called the standard model of particle physics which consists, roughly, of the following ideas: The world (and, more notably for my purposes, the biological and artifactual subdivisions of the world) consists entirely of a small set of particles (called quarks, leptons, and bosons), some of which (a subset of the quarks and leptons) combine to form the ubiquitous elemental atoms, which in turn combine to form the molecules of which all the substances and structures that we interact with are made. Interactions between any structures or patterns made of quarks and leptons are mediated by a set of only four fundamental forces (by way of the particles—the bosons—that are either known or theorized to mediate or "carry" those forces).[41] In short: in the vastness of the universe there are only particles and forces.

In addition, I assume that the behaviors of these material constituents are, for our purposes, deterministic. Nondeterministic quantum-mechanical effects are pervasive and can in fact be amplified to observable scales (see *e.g.* Chambers 1960; Tonomura *et al.* 1986; Young 1804); however, we have plenty of reason to believe that at practical biological scales (say, from the scale of biomolecular interactions up to the scale of blue whales or even ecosystems), those probabilistic effects tend to become statistically irrelevant—that is to say, they generally "average away"[42]. While biological systems could, in theory, evolve ways to take advantage of quantum indeterminacy, we don't yet have any good reason to believe that they do so widely, and there is neither any evidential

[41] The only force for which a mediating boson has not been found is gravitation. Many theorists assume the *graviton* will eventually be discovered. But we need not speculate here. Even if the mechanisms of gravitation turn out not to mirror the boson-mediated mechanism of the other forces, our story (of teleology and other material patterns) is about the material structures in our world and how forces act upon them. That story is independent of the behavior of bosons. What matters here is simply that there are forces that have effects on the motions of matter.

[42] For those who are interested, "statistically irrelevant" here means that the probabilistic effects average out as a result of the law of large numbers, so that the observable effects of populations of quantum phenomena fall in line with the predictions of classical mechanics. I am no expert in this matter, but please see Bohr (1976); Dirac (1933); Feynman (1942); and Tsang and Caves (2012) for pointers to Bohr's "Correspondence Principle", which is at the heart of this notion, and related discussions.

nor any conceptual link leading us to believe that goal-directedness or functioning (or for that matter, identity, or value) requires nondeterministic effects of any sort. The large molecules that comprise biological systems all appear to behave predictably and deterministically in their functional activities, as do the larger composite parts of organisms and their artifacts.[43,44]

This general doctrine about the substance and causality of the world is known variously as *naturalism*, *materialism*, or *physicalism*. It is important to mention because, as we saw to some extent in the previous chapter and as we will see to a much greater extent in the next chapter, human observations of purposeful phenomena were often explained by our forebears through appeals to supernatural, immaterial, or extraphysical phenomena. Even today, supernatural explanations continue to be assumed in some circles (*e.g.*, creationists). Moreover, as we'll see in Chapter V, the most popular modern theory of function (called the Selected Effects theory) is one that overtly claims to be entirely consistent with the modern scientific worldview, yet I'll show that this theory should be untenable to a materialist because its most widely championed versions implicitly entail either an extraphysical assumption or an anti-causal one.

[43] And if we can find a theory of these topics that is explanatorily and predictively powerful, while being independent of quantum effects, then so much the better for us in avoiding further headaches.

[44] Some modern artifacts (such as lasers and the atomic clocks used in GPS satellites) do depend on quantum effects in order to function properly. And, in principle, some biological items could do the same. But such things are out of the ordinary and so it is clear that while these quantum effects play a role *in a particular item's functioning*, they are unnecessary for the broader notion of *being functional.*

B. Thermodynamics

We grow in direct proportion to the amount of chaos we can sustain and dissipate.

—Ilya Prigogine and Isabelle Stengers (1984, p. 193)

Entropy is the price of structure.

—Ilya Prigogine and Isabelle Stengers (ibid, p. 283)

Good luck is rare, bad luck is the norm, and most problems left unattended don't improve spontaneously.

—Terrence Deacon (2013)[45]

In terms of my project, we'll find that not a lot depends upon whether or not some of the fine details of the standard model of particle physics are precise, or whether its parts might one day be further dissected or amended. What matters most of all is the gross picture it paints, with a presumption that, in one important sense, *nothing else exists*. That is, in order to make sense of the more complex topics of the special sciences, we must do so in terms only of phenomena that are *emergent from patterns* made of these fundamental point-like constituents and their physically causal interactions.[46]

[45] My understanding of the way the phenomena of thermodynamics relate to teleological topics ultimately differs from Deacon's (2013) treatment of these relationships, but there is no doubt that my view was influenced both by his and by Alicia Juarrero's (1999) work, which I recommend to the reader interested in further exploring similar ideas.

[46] See Holland (1998) for a good general introduction to emergence; and see the end of Chapter III of this manuscript for a few details about the history of the emergent perspective in biological thinking.

A well-known example of the way higher-level phenomena may emerge from patterns in the material world comes from the field of thermodynamics, the branch of physics that deals with the concepts of heat, work, entropy, spontaneity, reversibility, and order, all of which we will need to review since, together, they play a significant role in providing the context that allows for the emergence of teleology.

We're going to look at four main topics from the domain of thermodynamics: First, we'll look at the way properties such as temperature and pressure are emergent from collections of atoms. The fact that these new and fundamentally different *thermodynamic properties* arise from the probabilistic interactions of atoms and molecules will help make it plausible, via analogy, that new and fundamentally different *teleological properties*, such as value and goal-directedness, may also emerge from certain arrangements and interactions of matter.

Second, with the concept of kinetic energy available to us from the discussion about collections of atoms, we will examine the role that random or unpredictable kinetic energy plays in disorganizing the world. This tendency toward disorganization can be especially dangerous for living and teleological systems, which depend crucially upon maintenance of their organization in everything that they do.

Third, we'll review the famed *second law of thermodynamics,* which describes the probabilistic fact that a spontaneous reversal of disorganizing change is extremely unlikely. This view is often framed in terms of a direct synonymy between the notions of entropy and disorder, but I will follow recent work by Styer (2000) and Lambert (2002) in criticizing that comparison, and then I will offer a slightly modified basis for the concept of material disorder and of the irreversibility of that material disorder.

Fourth, we will look at a category of patterns—Prigogine's (1967) *dissipative structures*—that are in some way able to resist the irreversible tendency toward material disorganization. The

category of dissipative structures does not map directly onto that of teleological structures, but as we review these two categories we will come to see that the latter is a subset of the former. Dissipative structures form all of the material orderliness in our universe; they are the things that, by some method, can be constructed and maintained despite the tendency toward disorder. Teleological structures are one kind of dissipative structure. An important part of understanding what accounts for teleology will lie in discovering some reasoned method to further distinguish teleological structures from the non-teleological remainder of dissipative structures.

Emergence

In and before the eighteenth century, heat was thought to be a fluid called "caloric", which had a tendency to repel itself, thereby allegedly explaining the flow of heat from hot objects to cold ones. The intuition behind this theory is that if we can detect heat and if it can flow, then it must be a kind of stuff—it must be a material thing.

The truth, which took centuries of efforts by a great many scientists to discover and confirm, is that caloric does not exist. That is to say, the phenomenon of heat is real but there is no *substance* there, residing in or among the atoms making up objects. All that exists is those atoms and forces.

It was Daniel Bernoulli's (1738, 1741) innovative depiction of the interactions of those particles and forces that made possible our modern theory of heat, as well as of a number of related phenomena. Bernoulli was pondering the dynamics of fluids when he came upon a way to think of gases in terms only of the motions of a large number of atoms and molecules rushing about and colliding with one another. Although the idea was not confirmed or widely accepted until much

later[47], the form of Bernoulli's hypothesis was essentially correct: Each moving atom or molecule has a kinetic energy—or *energy of motion*—that is proportional to its mass and the square of its velocity[48]. Through collisions, this energy is constantly being exchanged with other atoms and molecules. It is continually being rearranged, and it escapes only when converted to other forms of energy (*e.g.* potential, rotational, or vibrational) or when transferred to the molecules that make up the container of the gas.

This image leads naturally to a number of conclusions. For one thing, the impacts of all these careening atoms upon the sides of a container should have the collective effect of pushing on the container; and the faster they are moving (or the heavier or more numerous they are), the stronger the total pushing should be (measured, say, in terms of a unit of area of the container's surface). In Bernoulli's picture, this pushing is what constitutes *pressure* in a fluid. Pressure is thus not a new or different kind of force, but is simply an outcome of the concerted motions of atoms in aggregate.

A similar aggregate quantity can be found in the average (translational) kinetic energy of the moving atoms, which, it has been determined, varies with the *temperature* of the gas. The faster the atoms move, on average, the hotter the gas, and the more slowly they move, the cooler. What we call temperature is a proportional measure of that average kinetic energy. Physicists have chosen to reserve the word "heat" to describe the amount of kinetic energy being spontaneously transferred

[47] After all, to truly believe the kinetic theory of gases, one must espouse some other underlying beliefs as well. For one thing, one must believe in the existence of atoms, a leap that wasn't widely taken until Einstein's (1905) explanation of microscopic Brownian motion in terms of atomic jostling. And for another thing, one must believe in the conservation of energy, which ensures that collisions between atoms are elastic, so that their motion continues until and unless the energy is transferred outside the system. See also Boltzmann (1872), Clausius (1857) and Maxwell (1860, 1873) for works further developing Bernoulli's kinetic theory.

[48] Curiously, Leibniz thought of this kinetic energy as constituting a "living force" that motivates behaviors, and so he called it just that, in Latin: *vis viva.* In the next chapter, we'll look at a number of historical concepts, similar to Leibniz's *vis viva*, meant to account for the vitality of life.

from one body to another, but nonetheless it is another thermodynamic phenomenon that arises from the aggregated microscopic motions of atoms[49].

To return, then, to our main point here: the phenomena of pressure, temperature and heat can be said to be *emergent* because they don't exist at the level of individual atoms, but only at the level of aggregations of interacting atoms. They are properties that emerge not from matter but from a particular arrangement of matter. And this is important to notice because the theory of teleology offered later will share this emergent structure. New properties—real, measurable features that are prominent parts of our world—including identity, value, goal-directedness, and functioning, will also be seen to emerge from a particular arrangement of matter. And no extraphysical substance or supernatural force—no caloric, no phlogiston, no *vis viva* or *élan vital*—will be required for that to happen. All that will be required is a particular form of organization. Just as is the case with pressure and temperature, it is the organization itself that will be shown to constitute the phenomena. Being purposeful, subjective, and evaluative will amount to nothing but a matter of pattern.

There's a long way to go before we can describe those teleological patterns. For now, we will continue looking at heat because it also plays a consequential role in creating the background conditions for teleology. Organisms and artifacts are all structures in which individual parts must hold certain causal relationships to one another in order for the whole to operate properly. Such items are orderly, and their orderliness is crucial to their teleological nature. If you significantly rearrange, damage, or remove parts, an organism or artifact will no longer be able to serve its former purposes. Because orderliness is so significant, the central context in terms of which organisms and

[49] The reason for this slightly unintuitive definition of heat is that different amounts of heat may be required to raise the temperatures of different objects by the same amount, depending on the *specific heat capacity* of the materials being heated. The details are of no great concern for the current work but, in short, not all the external kinetic energy that is added to an object by heat transfer is converted to kinetic motion within that object; some of it is converted or stored in other ways.

artifacts must be understood is their place in a universe that has a very general tendency toward disorder. We're going to analyze that tendency now, but it will take a number of steps to get where we're going, partly because the topic is a bit complicated but also partly because the prevailing modern view, given in terms of entropy, can be misleading. We can begin by noting the most important effect that heat has on organisms: they can be cooked.

Braising

The ancient foundations of the techniques that make up the culinary arts lie not in the pursuit of varied flavors to excite the palate, but instead in the pursuit of the simpler goal of food safety. As Bernoulli helped us come to understand, heating a thing only means adding energy to it by bombarding its surfaces with countless atomic and molecular collisions. It means violently, but microscopically, *shaking* the thing.

Over a century after Bernoulli's hypothesis, in 1862, Louis Pasteur demonstrated the truth of the germ theory—that life comes from life—with an experiment that showed that isolated, boiled broths would never grow new colonies of microorganisms.[50] In the wake of Pasteur's experiment, people came to realize that cooking with heat is actually a process by which various microorganisms—mostly bacteria, yeasts, and molds, but also a variety of parasites—become organizationally denatured and thus unable to replicate and grow anew. Today, what we understand

[50] Pasteur boiled solutions of broth in specially constructed swan-necked flasks that would allow the steam to escape, and air to be interchanged, but generally prevented dust particles from falling in. What he found is that normally occurring molds or bacteria would only ever grow in these sterilized (we now call them "Pasteurized") solutions after the flask was inverted or the swan-neck was broken, thereby allowing inoculation by microbes riding upon airborne dust particles.

this to mean is that the process of heating an organism will raise—eventually to certainty—the likelihood of catastrophically breaking apart the vital relationships that organize the organism[51].

Heat of course is only public enemy No. 1. To be sure, any source of energy that may impinge upon an organized structure has a chance of contributing to that structure's piecemeal disintegration, and so counts as an enemy. Electromagnetic radiation, for example, is another consistent source of flowing energy that can cause repeated and consistent damage to (even microscopic) physical structures, and this is why ultraviolet (UV) radiation has come to be used, like heat, as a modern industrial means of food sterilization[52]. And while heat and radiation produce cumulative microscopic damage, large energetic impacts, such as the damage done by a lion's teeth to a gazelle's flesh, or shoe soles to insects, or bullets to people, obviously count as disorganizing perturbations too.

The implication, when we understand that energy is distributed throughout the universe, is that every pattern that exists—you, me, and everything else—is being slow-cooked by the heat and radiation and the distribution of other energetic effects that occur in the world around us. Every material pattern in the world is being slowly shaken apart at a rate determined in part by its own structural integrity, and in part by the nature of the distribution of energy that impinges on it.

Despite the background of braising energy, there are two main reasons that you and I and other organisms are each here today. The first is that we (and our ancestors) have had a way to

---

[51] In military (and biological) strategy, *swarming* is a tactic meant to overcome a strong central target by using numerous decentralized (and usually much smaller) forces. Bees swarm over honey-thieves, armies swarm over cities, and so on. The idea is that the smaller and more numerous the assailants, the more *informationally complex* any defense against them becomes. If we take this idea to its logical conclusion, cooking (or heating in general) is using the smallest possible assailants—atoms or particles or electromagnetic waves—in the largest possible numbers in order to overwhelm even enemies that we are unable to see (such as bacteria). When enough heat or radiation is present, there is practically no viable defense (one can employ the crude low-information technique of *shielding*, but any shield has an energy limit beyond which it simply becomes inadequate). This is also of course, why heat and radiation are used not only in cooking but also in some of the most effective military weapons, in the form of firebombs and nuclear bombs. Although the use of overwhelming numbers of troops is in some way very different from the use of fire and radiation, there is a sense in which both are very much the same tactic of swarming.

[52] Not to mention microwave radiation used in microwave ovens for cooking, and X-ray and Gamma-ray radiation used in radiotherapy to break up the DNA of cancer cells.

counteract its damaging effects. We'll talk much more about that concept as we proceed. The second reason is that the earth is no longer a lump of molten rock with an atmosphere of steam, as it was during the Hadean period, over 4 billion years ago (Marchi *et al.* 2014). At the more moderate temperatures of most parts of our modern earth, damage to organisms is constantly occurring, but it can be considered relatively gentle.[53] We can exemplify the ubiquity of the problem by reminding ourselves of what is perhaps its most biologically salient result: The rather quick and inevitable desiccation of almost any organism's body that would occur if no additional water were being habitually mined from the environment to replace the water that is continually being lost.[54]

Environmental braising is certainly the major factor contributing to the deterioration of materially organized structures. Things might be quite a bit more resilient if they weren't regularly being struck by a distribution of damaging energetic impacts. But braising alone is not a guarantee that any particular thing will permanently fall apart. A second important contributor to disorder and destruction is the fact that most of this disorganizing change in our world is cumulative and generally irreversible—the random effects that pull things apart do not also randomly put them together (Achebe 1958; Yeats 1921). We'll return to the story of the irreversibility of braising and its effects on organized teleological patterns but before we do so, let's first review the thermodynamic subjects that historically have been understood to account for irreversibility. Not only do these subjects have their own important consequences for teleological patterns, which need to be understood, but studying them will also help to clarify irreversibility in braising.

---

[53] Astrobiologists have long considered planets that have moderate temperatures to be in the "habitable zone" of the stars they orbit, primarily because they allow the possibility of liquid water (Shapley 1953; Strughold 1953). More recently, the same idea has been renamed the "Goldilocks zone"—the region where things are neither too hot nor too cold. A more precise definition of the Goldilocks zone may one day be possible, given in terms not of temperature, but relying instead on, say, a ratio of the usable free energy required to do the work that can rebuild structural information with respect to the amount of structural information that can be expected to be lost per unit time given a particular distribution of energy flux through the volume of an open system.

[54] There are a few extremophile organisms that are able to survive desiccation and reanimate upon rehydration. But this is not an argument against the fact that braising occurs—they do in fact dry out. It just requires a different range of energy to further disorganize these organisms to the point that they can no longer be reanimated.

"Negentropy"

The notion of entropy, along with its tendency to increase—known as the second law of thermodynamics—has substantial implications for teleology and vitality. I would like to preface my introduction to these subjects by noting that the relationships here are more indirect than they are sometimes taken to be.

A classic example of expressing the relationships in an overly direct manner comes from Erwin Schrödinger who, in 1944, set much of the modern agenda for the branch of theoretical biology most closely related to these topics by claiming "a living organism . . . can only keep aloof from [death] . . . by continually drawing from its environment negative entropy" (or *negentropy* as Brillouin [1953] later coined it). Since then, more or less the same position has been repeated and cited often: entropy is disorder, vitality requires order, and therefore vitality is to be seen as consisting in processes of entropy reduction (or, as Schrödinger puts it, negative entropy consumption).

Once one understands entropy, Schrödinger's argument becomes intuitively appealing; however, it is only correct as a loose approximation. For one thing, one of the argument's premises is mistaken—entropy is related to, but it is not equivalent to, disorder in the material world (Lambert 2002; Styer 2000). And for another thing, even to the degree that one can accept the premises, the argument's conclusion is too simplistic—although order is necessary for vitality and teleology, that doesn't mean that any process that eliminates disorder or "consumes" order is sufficient for vitality (or teleology). In fact, many systems in our world (such as crystallization and the formation of stars) are able to either decrease entropy or to increase material order, or both, without displaying either vitalistic or teleological tendencies. (We will look at some of those systems shortly.)

Entropy and the Second Law of Thermodynamics

I am going to present entropy using the most popular introduction in modern pedagogy, even though that presentation easily leads one to mistakenly identify entropy with material disorder. There are three reasons for taking this misleading approach. First, the images used are clear and simple, making it easy to understand some of the basics about these rather abstract topics. Second, presenting things in this manner will bring to light the reasons behind the popular misconception regarding disorder, thereby allowing us to address and amend the error. And third, despite the drawbacks of these images, they nonetheless can facilitate, by analogy, an understanding of material disorder itself, given instead in terms of our earlier notion of environmental braising.

The traditional didactic method begins by describing what is called an *isolated system*—more or less a closed and thermally insulated box such as a thermos, with walls through which neither matter nor energy can enter or escape[55].

Our thermos is imagined to contain a gas composed of some moving atoms each of which has a mass and a velocity, and thus some kinetic energy. This all agrees with Bernoulli's picture and, as in that picture, the atoms are able to exchange energy with one another by colliding and thereby altering one another's velocities (their masses do not change; and, to simplify the image, we are assuming it is an inert gas in order to rule out chemical interactions).

In this picture, the quantity we call entropy is easily described, but difficult to get an intuitive feeling for. The Boltzmann interpretation says that entropy is a measure of the number of possible microscopic *arrangements* of the kinetic energy in the system that equivalently correspond to a single

[55] *Isolated* systems can be contrasted with what are called *closed* and *open* systems. In truth, there are no isolated systems other than the universe as a whole, since we can only approximately and temporarily prevent matter and energy from passing across boundaries. Closed systems are created more easily, as they only require the prevention of *material* exchange but allow *energetic* exchange across their boundaries. In general, however, any real (rather than idealized) province of the universe is an open system, where the boundaries are somehow specified, but both matter and energy may cross those boundaries.

macroscopic description that accounts for the distribution of energy in the same system. We can break that down to simplify it: A macroscopic description is one given simply in terms of the system's temperature, pressure, volume, and the number of atoms present. And the potential microscopic arrangements of a system are simply the different ways that all the energy in the system might be organized—some over here, some over there, some moving this way, some moving that way . . . So to make things concrete, a system consisting of one mole of helium atoms at a uniform 20º C and 1 atmosphere of pressure, within a thermos of volume 1 liter (altogether, these phrases constitute its macroscopic description), will be consistent with many different possible configurations of the ~6.0 x 10^{23} atoms within that space. The entropy of the system is a simple monotonic function of how many such equivalently consistent descriptions there are[56]. The diagrams usually used to clarify Boltzmann's description of *arrangements* are just flattened, two-dimensional pictures of certain states of a thermos full of atoms.

[56] To be precise, the Boltzmann-Planck formula for entropy, $S=k_B*ln(W)$, sets entropy (S) to be proportional (by way of Boltzmann's constant, k_B) to the natural logarithm of the number of possible equivalent states (W)—it is nothing more than a logarithmic transformation of a numeric count of those states (Planck 1901)

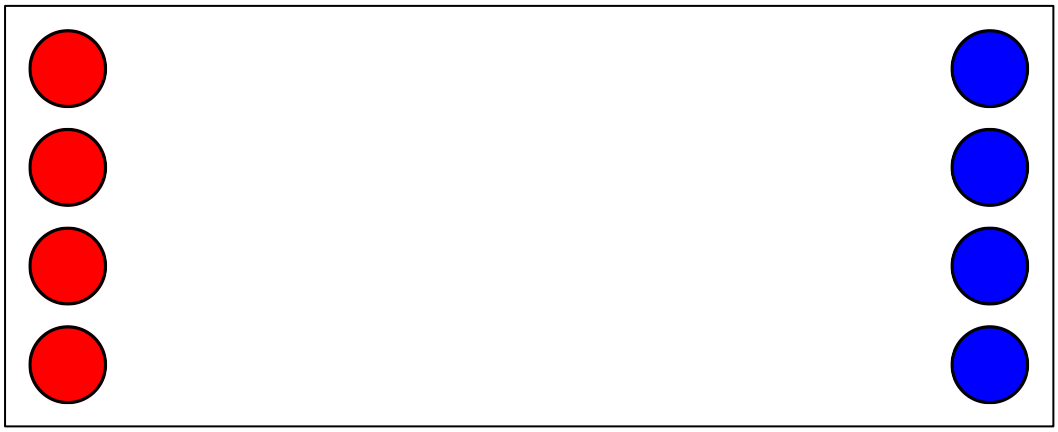


Figure 2.1.a: Fast-moving atoms lined up against the wall on the left (in red) and slowly- moving atoms lined up against the wall on the right (in blue). (We can assume for simplicity that all the atoms have the same mass.) It is very unlikely for a distribution of atomic energies in an isolated system to become naturally arranged in this macroscopically ordered way.

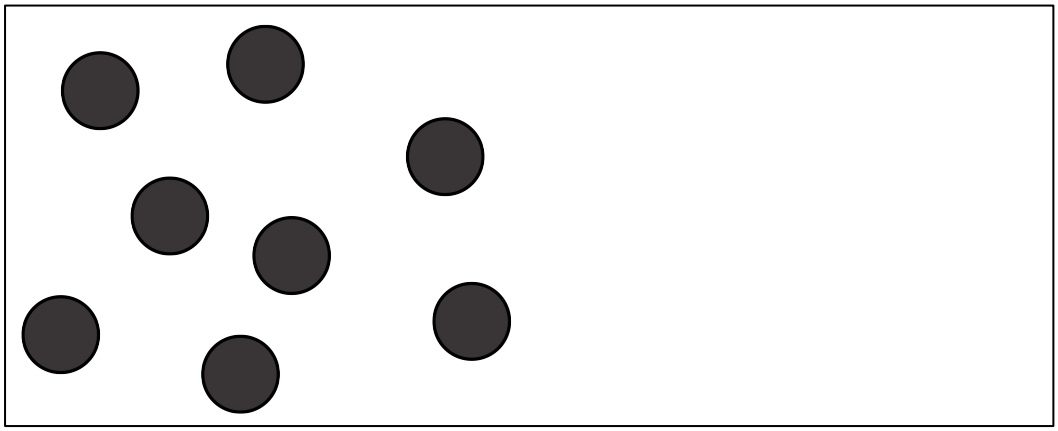


Figure 2.1.b: A group of atoms all located in one half of an isolated system. Because in this kind of diagram the atomic velocities and energies are unspecified, one might mistakenly assume that information to be irrelevant. Again, it is very unlikely for an isolated system to become naturally arranged in this macroscopically ordered way.

As shown in Figure 2.1.a, one can imagine that all of the *fast-moving* atoms (red circles) in an isolated system of gas happen to be on one side of the thermos, and all the *slowly-moving* atoms (blue circles) happen to be on the other side (let's assume for simplicity that they are all equal-mass

atoms). In this case, the entropy—the count of the potential states that have equivalently distributed kinetic energy—is relatively low. The reason it is low is that the scenario depicts a particularly well-organized arrangement for the energy in the thermos, and so there are relatively few ways to rearrange the energy within these atoms while maintaining the same description. (For instance, red circles could be swapped with other red circles, or blue with blue; but blue cannot be swapped with red without significantly redistributing the energy.) An alternative diagram, used commonly in thermodynamics courses and textbooks, is a black-and-white image, much like Figure 2.1.b, in which the distribution of velocities is unspecified (that is, there are no colors), but all of the atoms (or a significant majority) begin on just one side or in just one corner of the thermos (see, *e.g.*, Kondepudi and Prigogine 2015, p. 111). One can easily imagine (and thermodynamic analyses and models confirm) that, once the atoms are allowed to move, states such as those in both the images of Figure 2.1 become unlikely, while alternatives in which the energy in the atoms is further spread out across the whole space are much more likely at any moment in time.

In either case, the technical term used to describe both of the systems in Figure 2.1 is that the gas is *out of thermodynamic equilibrium*, which means that the energy distribution in the system is imbalanced. We can contrast this with being *in* or *at* thermodynamic equilibrium, which of course means that the system is in a kind of balance with respect to its distribution of kinetic energy. At equilibrium (and as long as the system remains isolated), because of the balance in energetic distribution, a system will no longer change spontaneously in either its macroscopic description or in the shape of the velocity distribution of its particles.

Now if time is allowed to flow, and thus the atoms in the out-of-equilibrium systems pictured above are allowed to move, they will interact with one another and with the walls of their respective thermoses, relocating and exchanging energy during each collision. Over time, the probabilistic result of this motion and these interactions is that the moving atoms in each box will

come to be mixed in their arrangement (rather than aligned against the two walls, or on just one side) and moderated in their speeds (fulfilling what is called a Maxwell-Boltzmann distribution, rather than the imagined bimodal distribution signified by the colors in Figure 2.1.a). All of this is to say that the system simply approaches equilibrium.

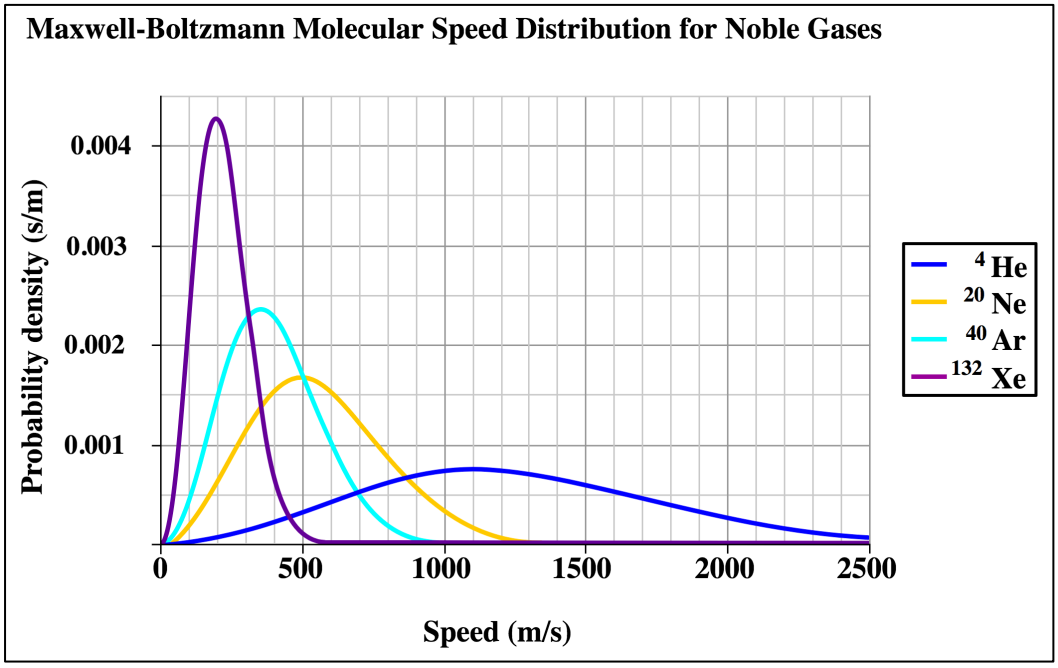


Figure 2.2: The Maxwell-Boltzmann speed distribution for four noble gases, at an equilibrium temperature of 25°C. Each distribution begins at zero (there are no negative speeds) and then forms a somewhat bell-shaped curve ending in a rather shallow, long tail (see Boltzmann 1872, 1877; and Maxwell 1860a, 1860b). The mean of each distribution varies inversely with the mass of the type of atoms it represents. When a system is out of equilibrium, the velocities of the atoms may fulfill a distribution of any shape but, as the system moves toward equilibrium, the shape of the curve shifts until it resembles a Maxwell-Boltzmann distribution.

Figure 2.3 shows the same two systems of gas from Figure 2.1 after they have reached equilibrium. In this figure, the entropy in each of the systems—the number of equivalent

distributions of the energy—is now at a maximum and the kinetic energy is distributed as widely and evenly as possible (Boltzmann 1896; Thomson 1852; see also Leff 1996).

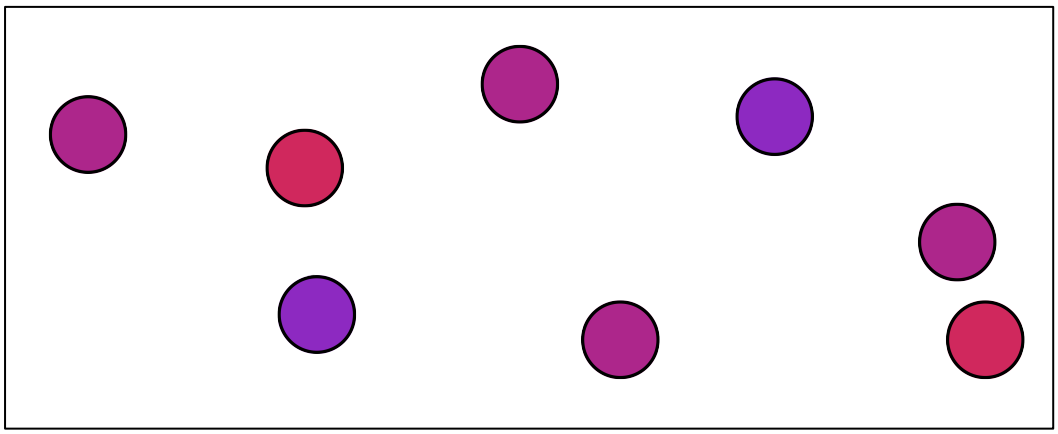


Figure 2.3.a: The same container as in Figure 2.1.a. However, here the atoms have all relocated and, through their collisions with one another, their distribution of velocities has taken on a bell-shaped but skewed distribution (see Figure 2.2). As before, the redder the atom, the faster it is moving, and the bluer, the more slowly; but now, as one can see, the velocities are clustered around a purplish mean, rather than around extreme reds and blues.

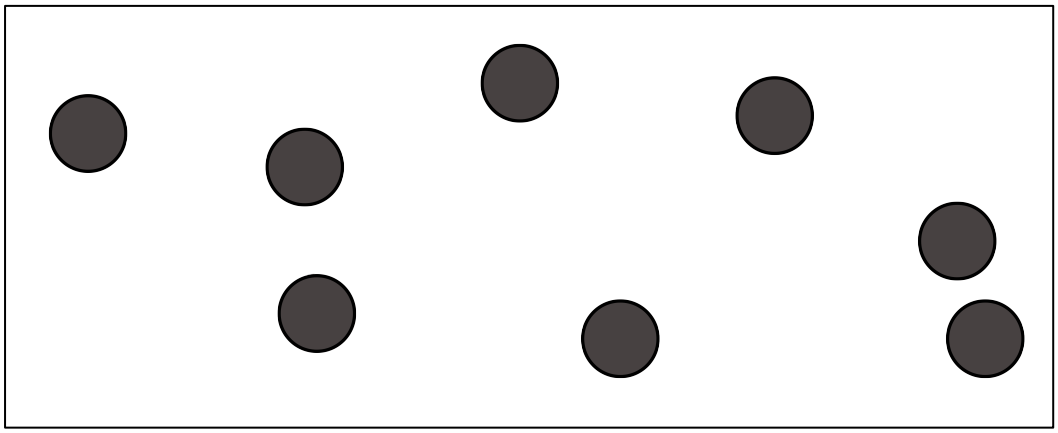


Figure 2.3.b: The same container as in Figure 2.1.b, with the atoms also having spread out over time through motion and collisions. However, in this illustration, as in Figure 2.1.b, the velocities and kinetic energies are still unspecified (and thus possibly assumed by a reader to be uniform or irrelevant).

At this point, one can intuitively understand that random rearrangements of the atoms will produce new and different microscopic arrangements, without significantly altering the distribution of energy in the system. And the fact of the matter is that, because of the statistical nature of large numbers (we assume many, many particles), any small alterations to the Maxwell-Boltzmann velocity distribution that are made by individual motions and collisions are always counteracted on average by other simultaneous or near-simultaneous motions and collisions, such that the overall distribution retains its shape (unless energy or matter enters or departs the system).

The thing that is particularly misleading about the comparison of the images in Figure 2.1 with those in Figure 2.3 is that one might focus primarily on the fact that, as time has passed, the atoms or molecules have moved. This can be even more misleading in the (colorless) b-versions of these illustrations that are so pedagogically common but that fail to specify any energy distributions or changes in velocities. What is most central to entropy and equilibrium is not that the *atoms* themselves have been moved or redistributed, but the fact that the *energy* in them has been, resulting in a more likely, more accessible, and more evenly distributed *state of energy*.

Entropy is not Disorder

Now that we have a sense of what entropy and the second law describe, we can see why they don't describe *material* disorder. In short, what they describe is *energetic* disorder, and these two kinds of disorder are correlated only to the degree that the material and the energy in a system are correlated in their arrangement. In a system where matter and energy are highly correlated—where energy disperses in large part through the translational motion of atoms[57]—entropy and material

[57] The paradigm case in which energy disperses primarily through the translational motion of atoms is the case when a gas in equilibrium is suddenly given more volume into which it may diffuse and expand. This is the case in the transition illustrated by Figures 2.1.b and 2.3.b, but it is a relatively infrequent occurrence.

order will also be highly correlated, but this isn't always or even often the case (see also Atkins 1984; Atkins and dePaula 2006; Lambert 2002; Styer 2000).

The misconception about disorder comes about when one relies on images such as Figures 2.1 and 2.3, especially in the black-and-white b-versions that do not specify the velocity distributions of the various atoms in the thermos. One might assume from such images that the energy in a system is able to disperse simply by being carried from place to place by the atoms as they move about. The mere translational motion of atoms, however, has no effect on their velocity distribution or on the number of equivalent microscopic arrangements *of energy*, to which entropy is tied; as was already described, the dispersal of energy occurs when it is transferred between atoms through their collisions. This latter mechanism can allow energy to move long distances while the underlying material substrate vibrates only locally.

Using the standard pedagogical thermodynamic imagery can be even more misleading, in fact, because the scenarios pictured are almost always imagined to be gases, and the correlation between the distributions of material and of energy is always much higher in a gas than it is in a liquid or, especially, a solid. And while the concept of material order applies to all phases of matter, it is most relevant in the solid phase, where orderliness tends to have the greatest stability over time.

Take, for instance, an elongated block of copper that is in contact with a heat source on one end, and a heat sink on the other end. The system is not isolated and not in equilibrium. But imagine now that we rapidly remove the heat source and the sink, enclose the block of copper in a vacuum-insulated box, and treat it suddenly as an isolated system. We can see, in such a case, that material order can be almost entirely decoupled from energetic order and entropy. The system begins far out of equilibrium, with energy distributed more on one side than the other, just as it is in the gases we've discussed. As time passes, the solid system is driven towards a uniform temperature and towards thermal equilibrium, by the transfer of vibrations in the atoms of copper. Ultimately,

the equilibrium entropy of the system winds up with a very different value from the entropy value at the start, but this change in energetic orderliness occurs in spite of the fact that the system will have changed little, if at all, in its material, structural orderliness. As long as the temperature is not high enough to melt the copper, the atoms remain for the most part just where they were, or very close to it.[58]

The block of copper shows that material orderliness does not have to decrease as energetic orderliness decreases, but there are even cases where material orderliness can *increase* during processes where energetic orderliness is decreasing (and entropy is increasing). For example: when a cloud of hydrogen atoms in a nebula is drawn together gravitationally to form a star, the conversion of gravitational potential energy into kinetic motion (that is: into heat) will increase the entropy of the system, as the particles move toward one another more and more quickly, and thereby jostle more and more violently. (And not only this, but if the atoms involved begin to move fast enough they will eventually fuse converting some of their nuclear potential energy into still more heat, increasing entropy still further). At the same time, though, as entropy (energetic disorder) increases, order (rather than disorder) in the material constitution will also increase, as the once widely distributed cloud becomes a localized, spherical star, and its constituents possibly fuse into heavier, more localized, stable elements.

Material order and thermodynamic entropy are phenomena that can be decoupled from one another. The theory of teleological patterns that I'll describe later is one based primarily on material organization—it is a theory of how certain very special kinds of material order come to exist and to remain in the world. The second law certainly applies here, but the tendency for energetic disorganization to accumulate is not the only thing that must be overcome in order for teleological

[58] There will of course be some expansion in the newly-warmed side of the copper block and some contraction in the newly-cooled side, but this is a minimal change in material locations of atoms, and an even more minimal change in terms of the material orderliness that is constrained by fairly rigid chemical bonding.

patterns (or, for that matter, other forms of material organization) to come to exist. The other thing that must independently be overcome, and that we'll look at presently, is the analogous tendency for material disorganization, caused by braising, to accumulate.

The Ratcheting of Material Disintegration

The spontaneous flow of events towards thermodynamic equilibrium is usually taken to be responsible for what is called the *arrow of time* or, in other words, the direction in which we perceive time to be flowing (Eddington 1928). The idea is that processes in which entropy comes to increase can proceed only in that one direction (of increasing entropy) and not in reverse, and so such processes are called *irreversible*. Heat flows from warm things to cool things, not vice versa. Systems such as those in Figure 2.1 evolve to look like those in Figure 2.3; but chaotic jumbles of whizzing atoms never suddenly end up in a state in which all the fast ones are lined up against one wall while all the slow ones are on the other side. When a dropped ball impacts the earth, its kinetic energy of falling is converted into heat as the atoms in both the ball and the earth set one another vibrating. Never, however, do a set of thermally vibrating atoms in the ground come to all concertedly push in the same direction at the same moment in such a way that a stationary ball suddenly rises up into the air. The only way to get such a reversal of normal events to occur is to perfectly coordinate a vast series of distant events beforehand. But barring acts of supreme omniscience, intelligence, and control, that kind of coordination is enormously unlikely to occur for even small numbers of particles. We can label this *the energetic coordination problem*. Certain energetic events can occur spontaneously in only one direction because, in order for them to occur in the other direction, it would require a hugely unlikely and complex coordination of prior energetic events. The second

law's guarantee is, in effect, a *ratchet* mechanism by which energetic disorder irreversibly accumulates (see also Deacon 2013).

But as I've said, in our quest to understand teleological patterns, we need to be interested in whether or not a pattern may maintain its particular form of material organization. As we just saw, energetic disorder and material disorder are distinct phenomena, and so a ratcheting effect in the one does not imply a ratcheting effect in the other. The isolated block of copper, for instance, is an example where energetic disorder ratchets up quite rapidly according to the second law, while material disorder changes little if at all. The second law does not imply ratcheted disintegration of material disorder.

We can, however, describe an analogue or a corollary to Eddington's arrow of time that does affect material disorder in much the same way. Quite intuitively, the random energetic impacts on a material pattern from an environment of braising may cause damage to the material structure that constitutes that pattern, and the next impact is more likely to further damage the pattern than it is to repair the damage recently done. We can illustrate this, as we did before with entropy, by considering a diagram of a boxful of atoms (see Figure 2.4).

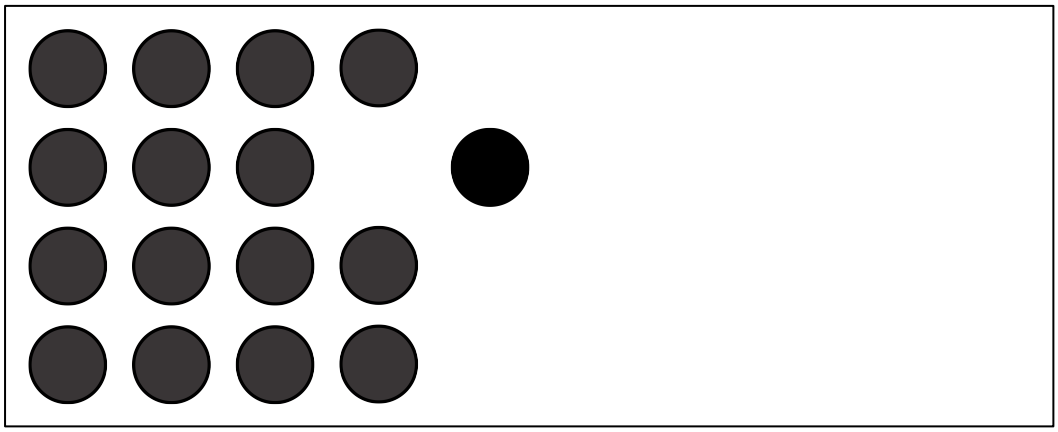


Figure 2.4: One atom, from amongst a set that were previously arranged in an orderly fashion, has begun to move. The picture is essentially the same as in Figure 2.1.b, except that material order has been emphasized pictorially. The atoms are all organized on one side of the available space, but their velocities are now truly unimportant to us. The atom that has ventured from the group now has exactly one way in which it can move in order to return to the orderly pattern, and a multitude of ways in which it might move further away from that particular state of orderliness. The same would hold for any of the other atoms if they also began to move from the arrangement. This illustration could be of either a fluid or a solid; the same considerations apply in any phase—as soon as any part of the structure is impacted by enough energy for it to break sufficiently far away from the rest, the likelihood increases that the lost part will continue drifting away, rather than returning to its previous location. "Sufficiently far" means that it has been knocked away to a location where the attractive forces pulling it back towards its previous location are overpowered by the other forces in its new location. In an environment with a distribution of energy that includes random perturbations that, on occasion, are able to move an atom to a distance greater than that threshold, the ratchet effect will ensure the steady and continuous disintegration of the material orderliness.

As one bit of matter—one particle or atom or molecule—is knocked away from the pattern in which it once took part, it immediately comes to have more ways in which it might move further away than ways in which it might go back. A familiar example might make this more intuitive: whenever you launch a washing machine through an asteroid belt and it starts bouncing off a

number of asteroids, the second asteroid strike is extremely unlikely to repair the damage made by the first; instead, damage to the washer accumulates and bit by bit it will eventually fall apart. The same goes, of course, for a washing machine that sits in a basement for six or seven decades. Although the damages sustained in the basement are more moderate than those caused by asteroid strikes, they also accumulate, and rarely if ever will they undo one another's effects.

Let's leave the washer aside now and talk instead about a dryer. This example will help to describe the irreversibility of material disorder and at the same time re-emphasize its distinctness from energetic disorder. If you take a bundle of folded laundry and put it in the dryer, it will all come out unfolded. If you take the same bundle of laundry, unfolded, and put it in the dryer, it will never come out folded. This process, in which unfolding occurs naturally and spontaneously but folding does not, looks very similar to the kinds of unidirectional examples that provide the basis for Eddington's arrow of time, but there is a difference. What is increasing and irreversible in these cases is not energetic disorder but material disorder. If we assume the two bundles of laundry to begin in the same energetic state (say, they are in thermal equilibrium with the room temperature) and if we assume they go through the same energetic process (a tumbling), then they will end up in the same energetic state: perhaps they will have gained a little heat and a little static charge, but in both bundles that change should be roughly the same, and thus irrelevant to the significant difference in how well folded the laundry is. In each of the bundles of laundry, the change in entropy over the course of the tumbling process will be equivalent; at the same time, however, the change in material order will differ significantly between the two cases.

From the washer, the dryer, and the atoms in Figure 2.4, we can see that material disorder, just like energetic disorder, has a tendency to accumulate in an environment of random changes. Said another way: material order has a tendency to disintegrate. This is an analogous, but distinct, arrow of time. We can call it *the ratcheting of material disintegration* or simply *ratcheted braising*. Under the

influence of random energetic perturbations, any material organization will have a tendency to fall apart. And ratcheted braising comes also with an analogous but distinct coordination problem that we can call *the material coordination problem*: The random background of braising is extremely unlikely to put together more than a few bits of structured information on its own and, even when it does, it is more likely to subsequently pull that structure apart again. In order for material organization to form, there needs to be some source of information or coordination that can drive the process. Just like a mechanical engineer's ratchet-and-pawl mechanism, the process of material disintegration is free to step forward, yet generally prevented from stepping back.

Framing the Context for Teleology

At last, we've reached our full description of the thermodynamic problem for teleology and vitality. It is not simply that entropy must be overcome, as Schrödinger had put it . . . although, at this point in our discussion, that problem remains too. In addition (and among other things), the material coordination problem needs to be solved in order to counteract the unstoppable effects of ratcheted braising, and to allow organized structures to come to exist and to persist.

Taken naïvely, these considerations seem to ensure gloomy prospects for our world. It seems as if nothing materially organized could ever take root in soil of this sort. But as we all know, that isn't actually the case. Organization abounds. Our world is full of fascinating patterns—from stars and planets and galaxies to diamonds and snowflakes and basalt columns, to gyres and storms, and bubbles and bedforms, and most especially, to living, teleological patterns. The world we live in is neither heaven nor hell; it is somewhere in between. It is full of ratcheted destruction, at every scale and at every moment . . . and yet at the same time, life blooms amidst the destruction; structure

and organization are created; and our world appears, somehow, to resist a complete descent back into the chaotic soup of particles where it began in the moments after the Big Bang.

There has to be a good reason behind the fact that order and life are able to bloom in a world where the cold statistical fact is that both energetic and material organization will tend to progressively disintegrate. Order can't arise by mere accident because, even if it did—even if a philosopher's "cosmic coincidence" were to create some organized structure in a small corner of the world for just a moment—that little accidental pattern would be subjected to as much ratcheted braising as anything else. It would quickly be torn apart. For ordered patterns to continue to exist in this world, something more than chance must be involved—some kind of *blueprint*, some dependable source of material orderliness must create them.

The theory of teleology described later is a theory of a particular kind of material orderliness. But in order to specify that particular breed of organization, we need first to account for the general existence of material orderliness. By and large, the thermodynamic principles involved in this have been worked out already. As we'll see next, much of the credit for our understanding of those principles is due to the work of the physical chemist and thermodynamicist Ilya Prigogine.

Death and Taxes

If we think of the second law and our analogous ratcheted braising as two universally imposed taxes—if we see them as inescapably taking constant bites out of our two orderliness accounts—then, in order to explain the organized patterns that we see, we'll have to look for some kind of workaround or exception to those universal statistical rules. Prigogine spent much of his career documenting and describing how this can happen. What we'll find, as we explore his ideas, is that there is indeed a pair of exceptions to the universal tax-codes, which, when coupled together,

allow us to get *Order out of Chaos* (as Prigogine and his colleague Isabelle Stengers titled their 1984 book). The first of these exceptions is a loophole in the second law by which certain systems can "move their accounts offshore" and thereby pay no "local taxes". Well known by physicists today, this loophole permits the energetic disorder of many systems to stay steady or even to decrease without violating the law. The second exception is a way to earn interest on material disorder before the ratcheted braising tax is applied and even, possibly, at a rate that may exceed the taxation rate. This gives systems that have low entropy (as created by the first loophole) the opportunity to reliably produce more material order, even while other simultaneous (braising) processes work to disintegrate it. It provides the tools by which certain corners of the universe may potentially come to be more and more orderly. We can look at the details of both circumventions now.

Far-from-Equilibrium Systems

The second law states quite simply that entropy will only ever increase. What this really means, though, is that it must always increase when measured over the entirety of the universe, or any other *isolated* system. In smaller and less idealized provinces of space, particularly in *open* systems where matter and energy both can be exchanged with the environment, the local entropy can be found to decrease, as long as the bottom line in our universal entropic bookkeeping works out to show an increase. That is to say, if the local decrease in entropy is simultaneously balanced by an equal or greater increase elsewhere in the universe, then the second law will not be violated.

As it turns out, all it usually takes for this to occur in any particular open system is for energy to flow through the system. The incoming energy, in whatever form it enters the system, will almost always represent an imbalance in the energy distribution of the system, and thus, in those cases, entropy will be instantaneously decreased, pushing the system away from equilibrium. If the flow of

external energy is consistent, then the system can be maintained in a state (or a series of states) that can be called *far from equilibrium* (Nicolis and Prigogine 1977). That comprises our first loophole—as long as there is a consistent flow of energy into an open system, there is a chance that the system can be maintained in an energetically ordered, far-from-equilibrium state.

One example of such a continuously energetically ordered (far-from-equilibrium) system would be the elongated block of copper we discussed earlier, with a heat source on one end and a heat sink on the opposite end. Another example would be the earth as it basks in the glow of the sun, always absorbing solar radiation on one side and emitting blackbody radiation (primarily in the infrared range) on the other. Another example is biological organisms, all of which absorb energy in the form of food (or sunlight, in the case of photosynthetic organisms) and thereby maintain an energy imbalance in the metabolic gradient between the absorption, storage, and usage of that energy. So far, this loophole accounts only for an imbalance of energy, or what we can call an *energetic potential*. It's a good start, but next we need to look at the further implications of that energetic potential.

Dissipative Structures

The second workaround to the orderliness tax laws is based on another pair of thermodynamic concepts that need to be introduced. As the entropy of a system increases and energy comes to be spread out, a quantity briefly mentioned earlier, called the *free energy* of the system, decreases (in ideal cases, reaching zero as entropy reaches its maximum). Free energy is the portion of the energy in a system that hasn't yet been evenly dispersed or distributed (Gibbs 1873; Helmholtz 1882). It is the fraction of the system's energy that is still out of balance. The reason it is important is that free energy is the only energy in a system that is available to perform what

physicists and chemists call *useful work*. As a system evolves toward equilibrium and its energy becomes evenly distributed, the free energy will inevitably dwindle, and so work, in any isolated system or any system into which energy no longer flows, must eventually stop. But in a system that is maintained far from equilibrium by a flow of external energy, some fraction of free energy will always be available, and so useful work can, in principle, always be done.

Work itself is of vital importance to our topic because it takes work to move things with mass, and material organization can only be established—structure can only be built—by moving matter into new locations. The workaround to the ratcheted braising tax law is that the free energy (created by the loophole in the second law of thermodynamics) can in principle be used to perform useful work that can create material organization.

Prigogine dubbed systems that take advantage of these two creative accounting techniques for accumulating earnings in spite of powerful taxes *dissipative structures*, because the processes that construct or maintain the order in them can only do so at the cost of generating entropy and dissipating that entropy along with some energy (primarily in the form of heat) into the environment. Dissipative structures exist in far-from-equilibrium systems that mine sources of energy from the environment, harness part of the free energy from within it to do (order-building) work, and return the unusable portion back to the environment in a state of increased entropy (see Prigogine 1967; Prigogine and Lefever 1968; Prigogine and Nicolis 1967; Prigogine and Stengers 1984).

The extraction of free energy from an environment is the basis for the biological concept of *metabolism* but, as we can see here, it is not an essentially biological notion; we've described it in purely physical terms. Other dissipative structures in far-from-equilibrium systems, such as growing crystals and the orderly convection cells that form in tropical cyclones, all have mechanisms that exploit energy flowing through the system, in order to perform useful, structure-building work.

We'll look at this distinction between biological and non-biological dissipative systems shortly, but first we need to address a subtlety here in terms of what the word "useful" means.

Usefulness

Usefulness is a concept that comes with slippery overtones of subjectivity and teleology. The term "useful work" sounds as if it is the kind of work that a person (or other agent) will benefit from. If we claim that some kinds of work are useful while others are not, but also that the useful kind can occur in crystal growth or storm formation, and so on, then are we claiming that these kinds of phenomena are already, teleological patterns? One would hope not.

I think a better term for the kind of work that can be performed by free energy and that cannot be performed in a system at equilibrium might be "usable work", which doesn't imply that someone actually benefits from that work whenever it occurs, but just that one could, if one so desired. I'll stick with the term "useful work" because it is already in widespread currency with physicists and chemists, but I'd like to further clarify its meaning in order to remove any teleological impressions that might accidentally arise.

For the purposes of our topic, I think the best way to understand this concept is to say that useful work is work that must be done while not simultaneously being undone. Any time two particles interact (through, say, a collision in a gas) they perform work upon one another. Each is made to move by the other. But at equilibrium, there is an important sense in which, on average, the work that is done through any particular interaction is also simultaneously undone by other particles interacting at the same time. Of course the individual changes—in those individual particles—are not undone, but the change to the overall distribution of energy is simultaneously undone. This is ensured by the statistical nature of equilibrium.

In addition, there is a second result at equilibrium. While there are material changes in an ever-moving gas (at every moment some particular particles will have gained some energy while others will have lost some) there can be no *concerted* material changes. The work performed by the transfers of kinetic energy in those many particles at equilibrium remains uncoordinated and thus unable to work together in a uniform direction. Because energetic disorder is at a maximum, the energetic exchanges through collisions are effectively random, and so the only effect this can have on material order is the disorganizing effect of braising.

However, when a system is out of equilibrium, potentially far from equilibrium, then the imbalanced (free) energy, in its attempts to become balanced as it drives the system toward equilibrium, can concertedly push in a single direction—from the imbalanced regions where energy is more highly concentrated, towards the regions where energy is less concentrated. When this happens, and the energetic flow in the system works concertedly, such organizational changes can occur as a gas pushing on a piston or gravitation pulling a nebular cloud of atoms towards the center of mass, or electrons flowing through wires to power electronic devices. And, if properly harnessed by material constraints (engine cylinder walls, wire insulation, and so on . . .), this capacity for material change can potentially be used to contribute to the production of material orderliness (see also Carnot, 1824; Deacon 2013).

Spontaneously Organizing Systems

Along with Gibbs (1873) and Helmholtz (1882), who each developed part of the notion of free energy, Prigogine has painted us a nice background picture of the thermodynamic requirements essential to producing the kind of material orderliness that underlies teleology and life. Teleological patterns can exist only in far-from-equilibrium, open systems because they are materially organized

patterns, and materially organized patterns can only come to exist and be maintained if there is a source of free energy that can be used to perform useful work in order to continually rebuild the material orderliness that ratcheted braising attempts to drain away. But these thermodynamic requirements are only *some* of the prerequisites for teleological patterns. And teleological patterns are only *some* of the materially organized patterns in our world. As we've noted along the way, the same thermodynamic processes also give rise to other patterns; free energy is also metabolized into the concerted useful work that organizes such non-teleological dissipative structures as stars and planets and crystals and storms and so on.

These other structures, and many more (a broader assortment of which are catalogued in Figure 2.5) form another category of phenomena that are often called *spontaneously organizing patterns*, because their organization seems to come out of nowhere, as a consequence of the laws of physics.[59] When the conditions are right, spontaneously organizing patterns simply materialize from their constituent parts.

[59] Although spontaneously organizing patterns are more often called self-organizing systems, I will avoid that phrase, as the word "self" has strong metaphysical implications, with regard to identity—implications that I think are more aptly applied to teleological patterns than to spontaneously organizing ones (see Chapter IX, pp. 435–7). Also, while most people might map the category of teleological patterns onto organisms, I prefer to remain open to the idea of teleological patterns that are not strictly cell-biological life.

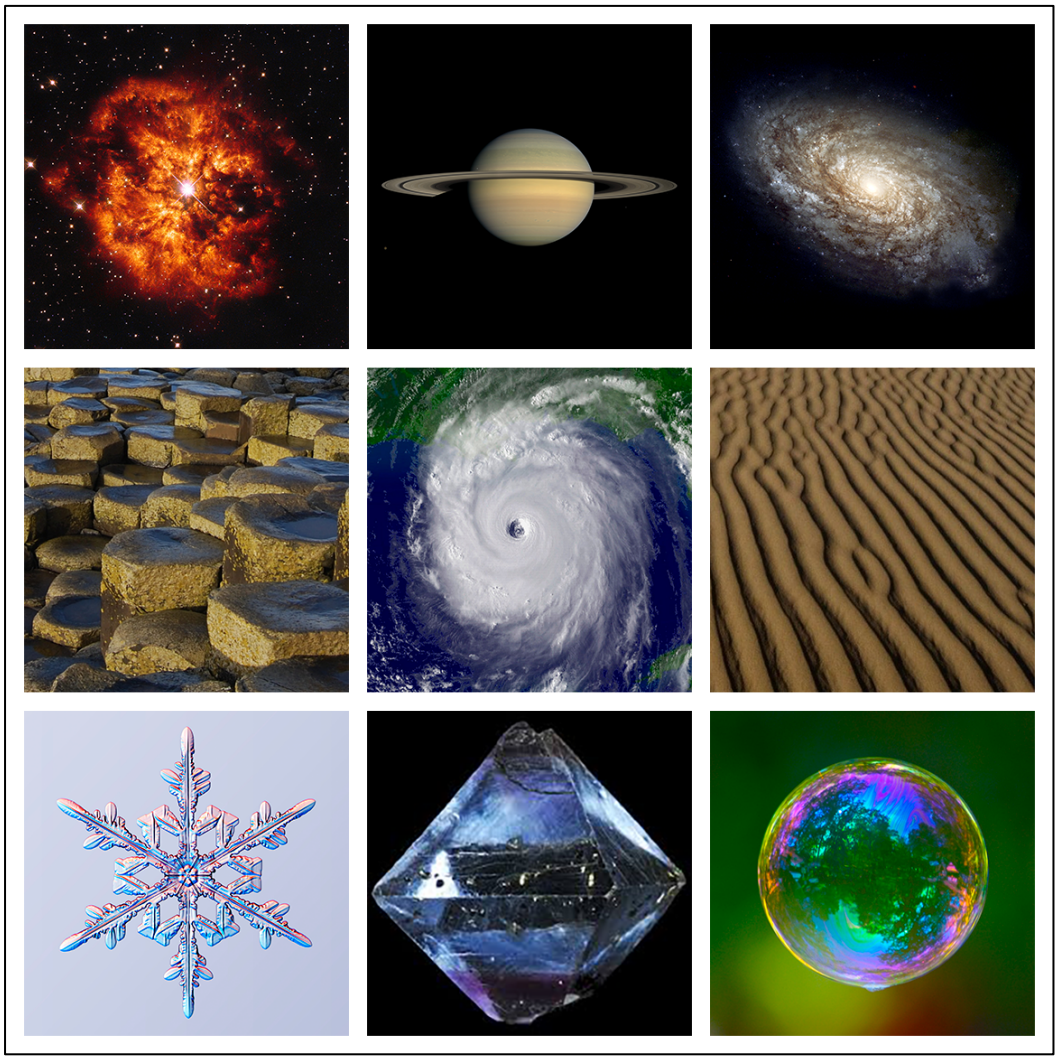


Figure 2.5: Various familiar forms of non-teleological material order, also known as spontaneously organizing patterns: (a) a star at the heart of a nebula; (b) the planet Saturn; (c) a spiral galaxy; (d) basalt columns; (e) a tropical storm; (f) a rippled bedform in desert sand; (g) a snowflake; (h) a fluorite crystal; (i) a soap bubble. (Nebula, Saturn, and galaxy photos courtesy of NASA. Basalt column photo taken by Petr Brož. Hurricane photo courtesy of NOAA. Snowflake photo taken by Kenneth Libbrecht. Fluorite crystal photo taken by Géry Parent. Bubble photo taken by Jeff Kubina.)

The features shared by teleological patterns and spontaneously organizing patterns have long caused philosophical difficulty for those working towards a definition of life. Suggestions about criteria for being alive have often included features such as growth, reproduction, and self-

maintenance, each of which only operates when coupled with a metabolism. But the main category of counterexample to this kind of definition typically has been various spontaneously organizing but clearly non-living patterns, examples of which are able to metabolize free energy that can then be used to reproduce (*e.g.* a candle flame), grow (*e.g.* a crystal) or to maintain their structure (*e.g.* storms).[60]

What Prigogine has given us is the thermodynamic principles that allow generally for material organization in the world. What have not yet been worked out are the principles that specify and guide the creation of that organization, as well as those principles that may allow discrimination between various kinds of material orderliness. A theory of teleology will need to distinguish those dissipative structures that are teleological from similar structures that are spontaneously organizing yet neither alive nor goal-directed.

We can make a short list now of some of the most obvious questions that Prigogine's paradigm leaves open for exploration: First, in any particular system, can enough material orderliness be produced to outstrip that which is lost to environmental braising? Second, if material order is being produced in a particular system, what would make that materially ordered structure goal-directed as opposed to spontaneously organizing? That is, are there characteristics of orderly patterns that can be used to differentiate between those that are merely orderly and those whose orderliness serves a purpose? And third, in either teleological or spontaneously organizing patterns, what might serve as the source of information or the blueprint that directs the construction of orderliness, thereby solving the material coordination problem? From where exactly do nature's blueprints come? There are almost certainly more questions than this, but I think being able to

[60] We will look later at the ways in which a theory that distinguishes teleological patterns from spontaneously organizing patterns may help specify a definition of life, and we can compare it to the type of definition mentioned here, which is put instead in terms of an essentialist list of criteria. For further discussions highlighting the difficulty of defining life in essentialist terms see, *e.g.*, Cleland and Chyba (2002); and the various other contributions to Bedau and Cleland (2010). See also Popa (2004), Trifonov, (2011), as well as the various responses to Trifonov collected in Volume 29, Issue 4 of the *Journal of Biomolecular Structure and Dynamics*.

answer these ones will already constitute a major step in the scientific explanation of material orderliness.

Cells, Cells, Cells, and 'celles

It will help us later in seeing how our theory makes the distinction between teleological patterns and spontaneously organizing patterns, if we have a few examples of each to hold up for comparison. Since we'll see a number of teleological patterns throughout the dissertation, at the moment I will just introduce some commonly discussed examples of spontaneously organizing patterns (see *e.g.* Deacon 2013; Juarrero 1999; Prigogine and Stengers 1984). But the goal here is more than just to familiarize ourselves with these examples; it is also to notice how these patterns resemble living systems in their growth, reproduction, and resistance to damage, while nonetheless lacking the hallmarks of teleology.

We'll look first at crystals (whose organization is based around a "unit cell"), then at storms (based on what is called a "convection cell"), then at another convection pattern that forms what are called "Bénard cells", and lastly at a class of non-biological structures called "micelles", which resemble biological cell-membranes both in their form and in their mode of assembly.[61] Crystals and micelles are both relatively solid structures, while storms and Bénard cells are more quickly evolving forms of organization that exist only within fluids.[62] We will begin by looking at all four types of pattern, but it is important to note that teleological patterns, such as organisms, never arise as purely

[61] It is a curious fact that most of the examples we'll look at seem to have been labeled as "cells", in one way or another. While the first three examples derive their name from Latin *cella*, which means storeroom or chamber, the etymological root of "micelle" derives separately, and only coincidentally, from Latin *mica*, which means crumb, with an added diminutive feminine plural ending (*-ella*).

[62] One category of spontaneously organizing patterns that we've already discussed—gravitationally formed celestial bodies—has both solid and fluid exemplars. Stars are composed largely of plasma; planets such as Jupiter and Saturn, of gas; and other objects, including the earth, the moon, and the dwarf planets Ceres and Pluto, are predominantly solid. There also happen to be planets, such as the recently discovered UCF-1.01, which circles a star designated GJ 436, that are made mostly of liquid magma (Stevenson *et al.* 2012).

fluid entities (such as patterns in convection), nor do they tend to occur as purely solid entities (as with crystals).[63]

By definition, a crystal is a set of particles (atoms or ions or molecules) that are organized through repetition at the microscopic scale. Each species of crystal is specified by its unit cell—the smallest volume of atoms or ions that, when tessellated in three dimensions, will construct the crystal. Crystals grow by the recurrent addition of the particles that form these unit cells, and those particles are usually moved into place by complex energetics, often including ionization from some source, chemical reaction, solvation, and then the thermodynamics by which ions or molecules come out of a supersaturated solution alongside the existing (seed) crystal and become more likely to bond to that seed than to re-dissolve.[64] Although the overall story is often rather protracted, in short, crystal growth exploits some of the free energy that flows through an open system to organize some of the system's material contents into a regular lattice (and we can be sure of this because crystals never form *at* equilibrium but always as part of the evolution of a dissipative system *towards* thermodynamic equilibrium). In addition to their seeding and growth, crystals may also at times split into parts (or "cleave") under stress, leaving each part to continue growing individually into distinct crystals in a process that thereby resembles reproduction.

Our next example involves the structure of storms such as tropical cyclones. A typical (non-cyclonic) storm cell is an organized structure that contains an updraft and a downdraft, together allowing circulation of air between the upper and lower atmosphere. The updraft that powers these

[63] We can speculate for now that one reason that organisms are generally a balance between liquid and solid structure is that the balance provides a two-part strategy that helps them resist braising. Unlike pure liquids, something with a partly solid component will be strong enough to resist the convective mixing and diffusion that might disorganize them. And unlike pure solids, something with a partly liquid component will be adaptable enough to be able to rapidly move resources to repair damages.

[64] For instance, the selenite (gypsum) megacrystals in the Cave of the Crystals in Naica, Mexico, are hypothesized to have been formed through a series of stages in which sulfide ions dissolved in magma-heated, calcium-rich water, and were then married with slowly diffused diatomic oxygen to produce SO_4^{2-} ions, after which both those ions and calcium ions slowly, over the course of a half million years, came out of supersaturated solution, and crystallized with hydration into $CaSO_4 \cdot 2H_2O$ (García-Ruiz *et al.* 2007).

cycles comes from the heating of air and evaporation of moisture over sun-warmed patches of ground or ocean. The convection produced in this process lifts the water vapor up to form a cumulonimbus cloud, which, when it reaches cooler high atmospheric temperatures, condenses and then falls as rain or hail. As the warm air rises, it leaves behind a low-pressure zone, which is then filled by drawing down cool air through the downdraft.

Tropical cyclones (such as hurricanes and typhoons) are even more structured. They form in roughly the same way, but some differences in their structure and conditions allow them to persist for quite long periods. When a group of single-cell thunderstorms form near one another and then converge over the ocean (as often happens near Cape Verde, for instance, where warm air coming off of the Sahara meets the cooler and moist coastal region of West Africa), the updrafts of those storms may merge together, resulting in a significant low-pressure zone. This depression attempts to refill partially by drawing downdrafts from above but also partially by drawing inflows of air from the surrounding surface region. Those lower-atmosphere and surface winds, end up spiraling inward and, in the process, becoming heated as they pass over the warm ocean. As the new air warms up, it fails to relieve the low-pressure zone that drew it inward, and instead comes to rise up to the upper atmosphere too, allowing the low-pressure zone to persist and to continue drawing in new air in a sustained process powered by the sea's energy.

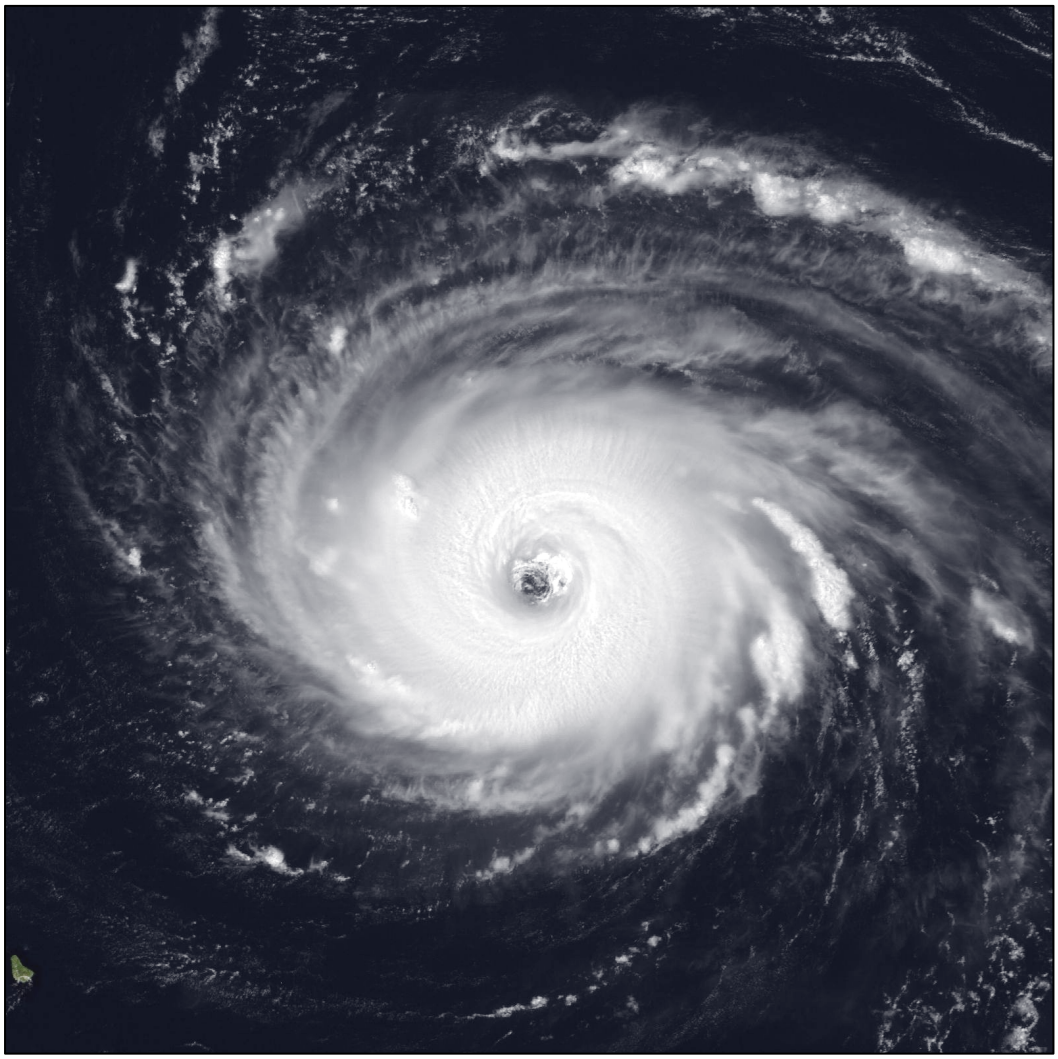


Figure 2.6: Hurricane Luis, a tropical storm traveling over the Atlantic Ocean, September 3rd, 1995. While the description in the main text focuses on the energetics involved in forming and maintaining a storm, an image such as this reminds us of the material orderliness that those processes produce. Photo reprinted courtesy of the U.S. National Oceanic and Atmospheric Administration.

In the case of cyclones, we have a process that avails itself of free energy (extracted from the non-equilibrium gradient between the warm ocean and the cool upper atmosphere) to maintain an organizational structure for a lengthy period of time. While crystals exhibit behaviors akin to growth and reproduction, the cyclonic process can, in some ways, be seen as akin to biological cell-maintenance because not only is the structure sustained, but also small fluctuations in the shape of

the storm (due, for instance, to collision with islands) get "repaired" spontaneously by the same energetic processes. Because of this, large tropical cyclones can persist for days or even weeks, unlike single-cell storms which usually disintegrate within about half an hour.[65]

Our third example is based in a process called Rayleigh-Bénard convection, which occurs when a thin layer of viscous liquid is heated from below in a gravitational field. At the start, thermal conductivity will transfer the heat energy, resulting in a linear temperature gradient from the bottom of the fluid to the top. At some point, however, as the heating continues, the fluid will rather suddenly and spontaneously begin to flow in an organized pattern, wherein the cooler, denser, fluid at the top will sink and the warmer, less dense, fluid from the bottom will rise in localized, self- and mutually-reinforcing regions called Bénard cells. The constant circulation within these convection cells is able to redistribute larger amounts of energy faster than conductive dynamics can.[66]

[65] An example of extreme cyclonic persistence is Jupiter's famous Great Red Spot, which has somewhat differing energetics but is also a kind of cyclone. The Great Red Spot has lasted at least as long as astronomers have been continuously observing it (which is over 185 years now, and possibly much longer). This longevity can be chalked up to the facts that the Great Red Spot has had a constant supply of energy, that it is kept more or less in place at low latitudes by interaction with other convection currents on the gas giant, and that surface friction in Jupiter's thick atmosphere is far lower than that which cyclones on earth's sea and land experience.

[66] See also the phenomenon of rose-window instability in a thin layer of low-conductivity fluid that is subjected to a high-voltage current (Niazi 2017; Pérez 1997).

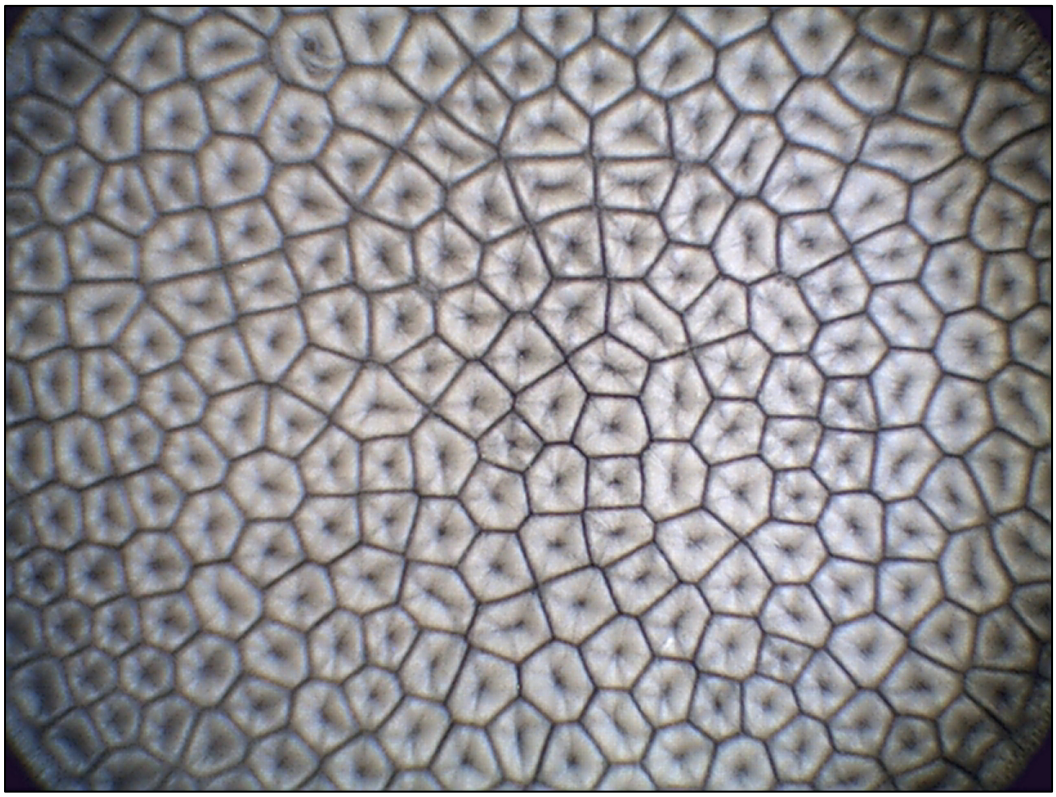


Figure 2.7: Bénard cells formed in silicone oil with powdered aluminum used as a tracer to create contrast. A fascinating fact about Bénard cells is that one can disturb the cell walls with, say, a toothpick or a spoon, and watch as they often reform more or less where they originally were. Photo reprinted courtesy of Vicente Perez Muñuzuri.

What is most surprising about Bénard cells is not that convection takes over because it transfers heat more efficiently than conduction—that much is an everyday observation—but rather that the form of these convection cells is so stable. They take shape with neat boundaries that are not permanent, but are nonetheless well defined and fairly persistent (see Figure 2.7). As we saw also with cyclonic convection, the biological activity that Rayleigh-Bénard convection most resembles in this regard is the metabolic process of cell maintenance. The boundaries of Bénard cells are able to resist moderate perturbations, quickly forming again after being nudged or disturbed with small instruments.

Fascinating as this phenomenon is, we need to be cautious in our analysis of it. The material order formed in Bénard cells is minimal. When a fluid made of a single type of molecule (or other homogeneous mixture) is made to move through convection, the overall *material* organization undergoes very little change, as each displaced molecule comes to be quickly replaced by another similar one. What is flowing in this circulation is primarily energy. That is not to say that there is no material organization formed, but the boundaries and centers of the cells that we can pick out observationally, differ from the rest of the fluid only in variations of density and surface tension, not by (semi-) rigid chemical bonding. The system is materially ordered, but neither highly nor durably so.

We can move on to our last example now. Micelles and their relatives (liposomes and phospholipid bilayers) are small structures that form by the spontaneous organization of certain molecules called surfactants, which are suspended in a liquid solvent (typically but not always water). Surfactants have the property of being amphiphilic, which simply means that they are elongated molecules possessing one end that is hydrophilic (attracted to water) and one end that is lipophilic (attracted to oils and not water).

When the energetic conditions are right[67], the hydrophilic ends of surfactants prefer to align with one another facing the water, so as to sequester the lipophilic (hydrophobic) ends together, in the now-protected interiors of the new structures they form. The same general plan is apparent not just in micelles, but also in liposomes and bilayers (see, *e.g.*, Bitounis, *et al.* 2012; Butt, Graf, and Kappl 2006).

[67] Micelles form only above what is called the critical micelle temperature (or Krafft temperature), which allows the surfactants to be freed up from their crystalized precipitate form, for other interactions. They also only form above the critical micelle concentration—a measure of how much surfactant is present in the solvent, and therefore of how likely the surfactant molecules are to run into one another in order for micelles and other such structures to grow.

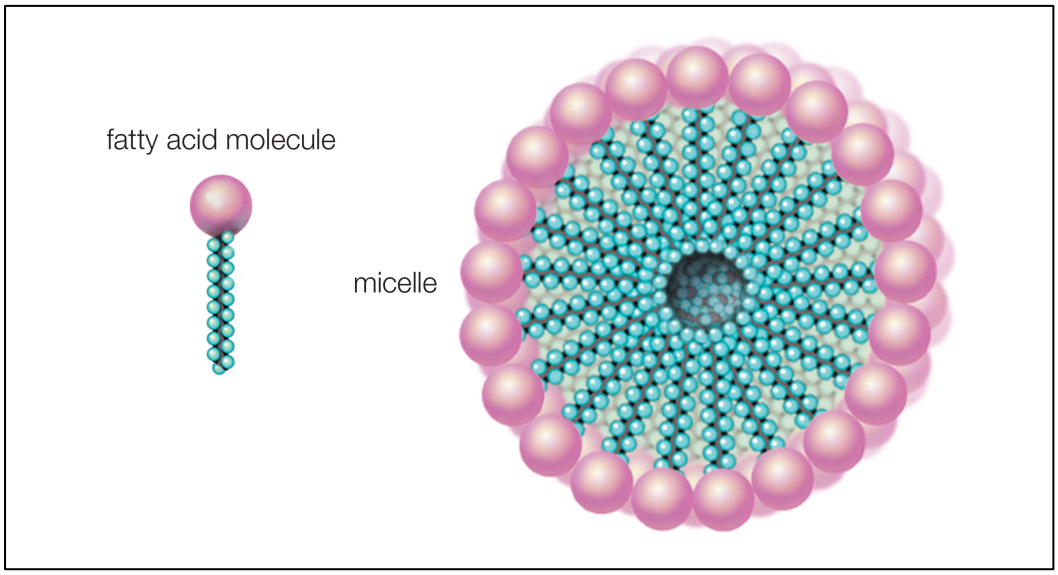


Figure 2.8: A schematic diagram of a micelle, composed of amphiphilic, fatty-acid molecules. The hydrophilic ends of these molecules are the pink balls in the diagram, while the lipophilic ends are the blue "tails" hiding in the interior of the micelle.

The processes of micelle formation and of crystal formation are quite different in some regards, but they share some thermodynamic aspects. In each, the phase change from a solution devoid of the ordered structures to one that contains them is always one in which changes in either concentration or temperature create an energetic imbalance in the system, which is then dispersed (to bring the system toward equilibrium) by the physical reorganization that accompanies the crystallization or micellization. Micelle formation also shares some properties with tropical cyclone development and Rayleigh-Bénard convection. As with those other two processes, the mechanisms that produce micelles are also able to repair them—as long as the conditions for formation remain the same, any ruptures to an existing micelle or liposome are soon filled with more surfactant molecules and thereby closed up. In micelles, we again see metabolic behaviors that appear to have some kinship to cell maintenance, by using energy and local molecular resources to reconstruct damage.

So as we've seen now, each one of these spontaneously organizing patterns results from a system that, in one manner or another, harnesses free energy from an out-of-equilibrium energy gradient to perform the work that coordinates the material organization of the pattern. Some of them grow and some of them replicate; all of them come to exist, all are able to persist, and all can heal some degree of damage brought about by external perturbations. Each seems to have some lifelike behaviors, and yet none of them has teleological tendencies.

Most notably, unlike agents (such as organisms) whose activities call for explanations in what-for terms, the things done by crystals, storms, Bénard cells, and micelles can all be accounted for solely in how-come terms, consisting of the physical proclivities of the systems that create them and the parts that constitute them. None of these patterns appear to be subjective or projective or to have strivings or goals of their own. Some of the patterns seem to have plasticity and perseverance but, for one thing, plasticity and perseverance are traits that aren't uniquely associated with teleology (think, again, of the pendulum or marble-in-bowl systems from the previous chapter) and, for another thing, the plasticity and perseverance in these spontaneously organizing systems is driven by the energetic processes of the whole system, rather than those of the material pattern that is created—it is a plasticity and perseverance that seems to be external to the orderly structure itself.

Schrödinger's Shadow

There is a sense in which Schrödinger's analysis of life in terms of negative entropy is correct: the second law is one constraint against which living things must struggle. As entropy comes to a maximum at thermodynamic equilibrium, the free energy of a system gets used up and so the ability to do any work that can maintain any ordered state diminishes. Without free energy, there

can only be an endless susceptibility to environmental braising, resulting in destruction and, eventually, certain death.

But there is also a sense in which Schrödinger's notion has constrained our thinking about what may account for life. By focusing on the notions of entropy and the second law, scientists following Schrödinger have missed the point that teleological and vitalistic patterns are, in large part, forms of material order. What it means for life to exist is not merely for patterns to struggle against energetic disorder, but more importantly for them to pursue environments of energetic order in order to extract a quotient of free energy that they can then use to do the concerted work it takes to struggle against material disorder. Life is more than just orderliness; it is a particular *kind* of orderliness, and if we want to understand that specific kind of orderliness, we need to step out from under Schrödinger's shadow and investigate the information-theoretic properties of material patterns, rather than those of heat.

C. Reality, Pattern, Organization, Causation

Information is information, not matter or energy. No materialism which does not admit this can survive at the present day.

—Norbert Wiener (1961, p. 132)

So I ask you, dear reader, are temperature and pressure real *things, or are they just* façons de parler*? Is a rainbow a real thing, or is it nonexistent?*

—Douglas Hofstadter (2007, p. 188)

Philosophers gather inquiries about what exists (or about what we can consider to be real) under the label "ontology", a topic that is not only a mainstay in the field of metaphysics but also deeply linked with teleology. In fact, the theory of teleology that we'll explore in Part II constitutes a particularly profound ontological claim. It rests upon a distinction between three major categories of patterns that can exist within our universe, two of which I have already named: spontaneously organizing patterns and teleological patterns. The third major category consists of patterns that, because of incessant universal braising, become only less organized, never more so, over time. I'll call patterns in this third category *ontological nonce*, or just nonce.[68] Teleological patterns and spontaneously organizing patterns are structured forms of material orderliness. They are patterns

[68] I have borrowed the word "nonce" from lexicography and cryptography where it refers to a word or a token of data that is intended to be used on just one occasion and then discarded. The analogy here is that the patterns I call ontological nonce are also expected to exist only momentarily, and then bound to become something else. Curiously, the term "nonce" has its etymological roots in the Middle English for "the one purpose" (*then anes*). The word attracts me more, however, because it *looks* as if it is a portmanteau of "not" or "nothing" and "once", thereby impressionistically reflecting my notion: patterns of ontological nonce can be considered to be not things ("nothings"), or things that exist unreliably, briefly, . . . or, as it were, just "once".

brought about by reliable methods for creating order; blueprints for their creation and maintenance are distributed, somehow, within the world. In contrast, items of ontological nonce might in some sense also count as patterns—they may be perceptible and thus, for a while, identifiable; but unlike the members of the other two categories, the material organization within them is fleeting and unreliable, and there are no blueprints anywhere that could contribute to their maintenance.[69]

One of the key strengths of this tripartite theoretical distinction lies in the fact that it is not merely a *philosophically* ontological claim—that is, it is not metaphysics. What I will present later is, rather, a *mathematically* ontological claim, a formulaic way to partition the space of all potential patterns. The distinction produced by that manner of partitioning arises naturally, given a small set of (what I think are) physically justifiable assumptions about the material world. Among those assumptions are, of course, the notions of material orderliness and braising that we've just explored, but also the basic ontological proposition that I'll try to argue for now: the claim that, at least in some reliable, overarching sense, our world is made principally of patterns.

Some readers may already be inclined to see the world as pattern, or as information, or as mathematics, or may already have some other interpretation of reality that lies roughly along those lines, and if you are one of them, then what follows should be easy to swallow. Such interpretations have certainly been prominently argued for before (Dennett 1991; Gershenson 2010; Hofstadter 2007; Tegmark 2014).[70] As we'll see in a moment, a notion of this sort is also implicitly entailed by any doctrine that firmly marries the modern scientific worldview with the emergent perspective. I

[69] Items of nonce may seem to persist for quite a while, relative to subjective human scales of time measurement; however, that persistence depends only upon their relationship to the local distribution of braising energy.

[70] In his 2014 book, *Our Mathematical Universe*, Tegmark follows Eugene Wigner (1960) in asking why mathematics is so strikingly effective at describing the world (see also Holland, 1998, who speaks of the "unreasonable effectiveness" of mathematics). The answer that Tegmark gives is that the world *is* mathematics, in some strong sense of the term "is" that he leaves not fully explained. This may seem a radical view to some but, in certain respects, I don't think it differs too radically from the view I am giving here. My assumption, based on the modern scientific worldview, is that the world consists of some basic parts, which, thanks to their regularities, may combine into untold myriads of patterns, all of which can at least be mathematically or informationally *described* (see, *e.g.,* Galileo 1623). And, at some point, when inquiring into the difference between "being" and "being describable as", one may find oneself only a mild semantic stumble away from Tegmark's view.

happen to think this latter conception is the clearest way to understand the world as consisting of patterns.

The Emergent Perspective

Here, I will follow Hofstadter (2007), who, in developing his claims that selves or souls are real things (and his ensuing arguments for just what kind of real things they are), bases his underlying argument for the-world-as-patterns in the notions of causation and emergence.

Hofstadter frames his discussion in terms of "the causal potency of patterns", by which he means that anything that is a pattern deserves distinct consideration as a thing of its own kind because the pattern itself has new, reliable consequences (causal potencies) over and above those distinct underlying consequences of the atoms and particles of which it is made. To illustrate, he highlights "that evolution caused hearts to evolve, that religious dogmas have caused wars, that nostalgia inspired Chopin to write a certain étude, that intense professional jealousy has caused the writing of many a nasty book review, and so forth and so on." Hofstadter points out that the straightforward resolution to the tension that seems to exist between these higher-level kinds of causation and the physicist's lower-level conception of causation in terms of the four fundamental interactions, lies in understanding that "these 'macroscopic forces' [are] merely *ways of describing* complex patterns engendered by basic physical forces". And he reminds us that this shouldn't be a difficult leap to take, since physicists have already shown that many other macroscopic forces and phenomena, such as "friction, viscosity, translucency, pressure, and temperature can be understood to be patterns made of interacting particles and their forces" (*ibid*, p. 33).

Shape and Causal Dynamics

We can perhaps briefly expand on Hofstadter's way of putting things. A physical pattern of interacting particles and their forces—an atom, a molecule, a bone, a hammer—is also distinct from those underlying particles because of the ways in which the correlations between the particles transform the potential interactions they may have with other patterns. What otherwise would have been random (uncorrelated) interactions become new kinds of concerted joint activity. The pattern itself—the organization among the particles—allows for behaviors different than those that the same particles would perform if they were not so correlated. Roger Sperry (who, incidentally, also influenced Hofstadter's view on these topics) gives an example of "a wheel rolling downhill [which] carries its atoms and molecules through a course in time and space and to a fate determined by the overall system properties of *the wheel as a whole* and regardless of the inclination of the individual atoms and molecules" (1980, p. 201; see also Sperry 1969). The atoms in the wheel have no individual tendency to take the spiraling courses that they do, yet when arranged in the mutually constraining form of the whole wheel, they have no choice but to follow those paths. It is the organization of the wheel that matters.

To be more explicit, we can recall that physicists and chemists think of particles as resembling tiny point-like magnets that exhibit four very different types of "magnetism" (the four fundamental forces). All of the world's particles simultaneously push and pull on one another in these four independent ways. And because of the complex interplay between the various forces in various kinds of particles and the varying strengths of those forces at varying ranges, the world comes to be filled with patterns of myriad kinds, some of which "like" to bind closely to one another, others of which like to push each other away, and others which just seem to be indifferent to one another (to speak teleologically about things that are not at all teleological). Every pattern is

just a combination of these little "magnets" whose fields of pushing and pulling deform and combine to produce new, unique shapes with new, unique tendencies. Chemistry, of course, gives us the clearest, simplest examples: every type of ion or molecule has a specific shape that gives it characteristic—and often well-documented by now—behaviors with respect to other particular patterns. But complex aggregates of particles that go beyond basic chemistry—macromolecules, polymers, everyday objects, and even large gravitationally bound coalitions—all have their individual capacities to affect other patterns, too. Each pattern does what it does because the organization within its set of particles collaboratively produces its distinctively shaped causal dynamics that it then imposes upon the patterns around it.

Four Kinds of Causation

One thing that makes this view of patterns appealing is that the total picture that physicists have drawn for us turns out to account for three of the four classical Aristotelian "causes"—the material, formal, and efficient causes[71]. All of our "how-come" explanations for why things come to be can be given in these terms. When we explain something in terms of its material causes, what we are really talking about is the micro-dynamics created by the organization of the particles that make up the material. The causal dynamics created by the shape of a thing at a very small scale explain, in part, why a thing is what it is or does what it does. When we explain something in terms of its formal causes, we are talking about the macro-dynamics of the thing. The causal dynamics of the overall shape of a thing explains another part of why a thing is what it is or does what it does. And when we explain something in terms of its efficient causes, we are talking about the historical series of incidents in which the initial movements of various patterns, combined with the possible

[71] We'll explore the philosophical history of this classification in Chapter III.

interactions of those patterns by way of their material and formal causes, have come to result in some event or state of affairs (or the existence of some new pattern).

The only "why" that is unaccounted for in all of this is the final (teleological) cause—the explanation of a thing in "what for" terms (see Chapter IV for a dissection of the word "for"). Teleological causes are not present in all events or objects in our world but, in those cases where they are present, they need to be accounted for. As it turns out, viewing the world as consisting of dynamical patterns is the first step towards explaining the emergence of teleological causation. The theory we will explore in Part II claims that a certain category of potential interactions between those patterns is what differentiates teleological organization from either spontaneous organization or lack of organization.

The Summary of These Parts

The perspective I've been advocating, from which we might see the world as being made of particles but populated by *patterns* of those particles, is not meant to be a complete accounting of the "furniture of the world" (a term philosophers often use to describe their theories of what exists). Although it may one day be possible to more rigorously defend some more careful version of this account, at the moment I am not working towards such a defense but only setting out the assumptions that my primary theoretical work will later rely upon.

The account is incomplete for a number of reasons. For one thing, I have been discussing only patterns that seem to be materially organized by the bonds within them, not patterns that seem to be organized primarily in time (as in sound waves), or organized in space, yet only correlated by incidental history rather than mutual bonding (as with the photons that make up the patterns we perceive as rainbows). For another thing, the account does not address patterns created by a

process, such as the rippled bedforms seen in Figure 2.5.f. For a third thing, the account does not address abstract or distributed "objects"—such as governments or satellite communication networks—as patterns. I tend to believe that many of these things (those that are not illusions) may be patterns in the very same sense[72]; but I am not yet prepared to carefully defend that position, and I prefer to postpone the depths of those analyses.

At the moment, all we really need is a compelling argument that *many* things in our world are patterns of this sort, and that those many types of patterns may interact to cause changes in one another. With that assumption to stand on, we can then begin to build the machinery (in Part II) that can be used to explore whether or not that assumption, along with some others, is enough to give rise to patterns or systems that behave teleologically. I think it is.

There is, however, a more concerning problem that I see with this account of patterns. It seems to me that if we claim that the world is made of correlated groups of atoms, we will be plagued with the problem of identity. Since every particle's fields of force extend continuously and infinitely, how exactly should we carve the world into those groups, saying which parts count as being correlated and which do not? How do we say that this proton and electron form an atom, but another nearby lone electron is not part of the same atom? I'll introduce the problem of identity in more depth after we address another ontological topic—that of patterns that are just illusions—but I won't try to solve it until much later. The answer as to which versions of a pattern count as being "that pattern" will turn out to be something very akin to the common notion of functional equivalence. I will call it causal equivalence, because I would like to avoid using the term "functional" unless a pattern is involved in a teleological system. That, however, is only a minor quibble over words (see also Chapters IV and V).

[72] I am convinced both by Hofstadter's lengthy argument by examples (as cited above and also drawn out in more detail in his 2007 book) and by the obvious possibility of long-distance correlations (potentially mediated by coordinating elements) that may account for the joint action of seemingly disjoint parts. Communication, for instance, allows for the coordination of the motions of distant objects.

D. Illusion

The human understanding is like a false mirror, which, receiving rays irregularly, distorts and discolors the nature of things by mingling its own nature with it.

—Francis Bacon (1620)

When we look at the topic of function in Chapter IV, we'll find that there is a special pattern that many scientists and philosophers come to see and that, in seeing it, they come to take it to be a legitimate phenomenon of their study—a pattern that is really out there, in the world. The particular pattern I am talking about here goes by the name "proper function", a term that refers to the idea that an item (for example, a heart or an eye or a pen or a chair) might have a duty to carry out regardless of context, or of damage, or even of a failure to have been properly produced. A proper function of an item is a property that is imagined to have been bestowed upon the item and then, somehow, to reside metaphysically within it.

When we come to that discussion, I will dispute the claim that functions are properties of this sort, claiming instead that this commonly observed pattern is only the result of a shadow cast upon the mind with such regularity that we find it hard to believe that it is not a constant part of the world. I don't intend to dispute that items may function; what I dispute is that they *have* functions.

A bit later, in Chapter VII, we'll find that there is another pattern that a preponderance of modern scientists and philosophers admit seeing, and yet which they take to be an illegitimate phenomenon not worthy of serious study—a purported illusion that they refuse to believe veridically reflects any actual facet of the reality we live in. This pattern—goal-directedness—is in fact one of the central topics of study in this dissertation.

In that case, my disagreement will travel in the opposite direction. I will claim that goal-directedness is very much an objective pattern that occurs in our world, and that we have no convincing basis—not a shred of evidence—upon which to base any disbelief in it.

Because of these two very fundamental disagreements that I hold with a large number of scientists and philosophers about the nature of the subject matter of teleology, I must spend a few pages now discussing the notion of illusion, and the criteria by which we ought to recognize some perceived patterns as being real, while discounting the objective existence of others. I hope to remind us of the tools we have available, used regularly both in everyday reasoning and in the scientific method, to determine which of the patterns in our perception reflect the outside world faithfully and which do not.

Color Constancy

One widespread although not too troublesome illusion is the notion that things have colors—for instance, that the sky is blue or grass is green or apples are red. For some the illusion is pervasive: to children, that the sky is blue seems incontrovertibly true. Even for the scientifically educated who know well that things change colors under varying circumstances (and that the vibrant and varied qualities of coloredness only exist in the minds of perceptual creatures), the illusion of color constancy is still reflected in our everyday ways of talking about colored things.

Illusions such as color constancy are often called "useful fictions" because, while the impressions they give us about the world are false, in general it doesn't hurt to talk about them as if they were true.[73] For most practical purposes that relate for instance to apples, speaking of them as

[73] The most pervasive illusion of all is the indispensable fiction whereby our internal perceptual experiences of each facet of the world are collectively brought together to appear as if they simply *are* the outside world.

if redness were an intrinsic property won't cause any confusion and it will actually help us greatly in categorizing and collecting apples.

However, at some point in scientific inquiry our needs change and we want to understand a phenomenon in all the ways it manifests, including not just its common and central cases but also its exceptional or boundary cases too. At some point for instance we no longer want to just collect apples; we also want to understand both the nature of light and that of vision (or, in my case, the nature of function and that of goal-directedness). And when we reach that point in our inquiry, we need to carefully work through which of our perceptions about the phenomena related to our topic are real and which are illusory, eventually replacing, for instance, the idea that "things have colors" with the idea that "things reliably appear colored" (due to their, and our, reliable interactions with light).

Determining Illusion

So how do we know when a pattern that we are observing is an illusion? Well, obviously we can't always know; that's the nature of illusion. But there can be some telltale signs, and I think looking at a few more examples of well-known illusions will help us remember how to look for those indicators.

As most of us are aware, the image of a ball moving across a movie screen is not a ball at all, and neither is it even an image of a ball moving across the screen. There are actually two simultaneous illusions. One is the cinematic illusion of motion, created by way of a series of images presented in quick succession at different locations. And the other is what we can call Magritte's illusion: there is no ball on the screen just as there is no pipe on Magritte's canvas—there is only the reflection of light onto our retina in a pattern similar enough to that made by an actual ball or pipe.

Figure 2.9: *The Treachery of Images*, a 1929 oil-on-canvas painting by René Magritte, © 2018 C. Herscovici / Artists Rights Society (ARS), New York. The caption translates to "This is not a pipe", which is of course true. Magritte's point—Magritte's illusion—is that, in the case of paintings and photographs and so on, our minds have a tendency to reconstruct the whole of the three-dimensional objects depicted.

One way for the naïve moviegoer[74] to determine that a cinematic ball is not real would be to interfere with the screen's ability to reflect an image, for instance by shaking it, or viewing it from the side, or coating it with Vantablack™.[75] Another way would be to throw another (real) ball at the first one; we would find the two balls wouldn't interact the way we normally would expect. A third

[74] For an early philosophical exploration of cinematic illusion, see Plato's "Allegory of the Cave" (*The Republic*, 514a-520a). In the story, Socrates describes a group of people chained inside a dark cave where their only perception of the world is through a puppet show projected upon on the wall by firelight. In order to prevent the individuals in the cave from discovering the illusion, Socrates designs the story such that any chance for manipulation is wholly (even if unrealistically) restricted.

[75] Vantablack™ is currently the world's blackest artificial substance. It absorbs more than 99.9% of visible spectrum light, reflecting almost none. The material has many potential uses, but would make for the worst possible movie screen.

way would be to shine a bright white light at the ball to try to highlight it; the actual effect will be to drown out the image. A fourth way would be to discover the projector and the filmstrip, and to interfere in any of a number of ways with how the light is cast onto the screen. Any of these little interactive experiments would expose the fact that there is no ball—none of the predictable causal effects of "ballishness" take place.[76] And yet if there really was a ball, the same experiments would inform us somewhat by neither destroying the ball nor interfering with our perception of it. (For more on predictive success as a diagnostic of the reality of perceived patterns, see Dennett 1991; but for complicating cases, see the above discussion on useful fictions, such as color constancy.)

Have a look at the Hermann grid illusion shown below in Figure 2.10. One way to notice that the intersections in this image don't really harbor little glowing circles is to try to train one's eyes upon one of those circles. When we do, the one we are trying to look at disappears. They all flicker in and out of existence as a function of how directly or peripherally we try to look at them.

[76] To put it in more scientific terms: The hypothesis is that there is a ball. The prediction, from that hypothesis, is that manipulations such as painting the screen or shining a light at it should change little or nothing about the ball (hitting it, however, might change its trajectory). If any of those predictions turns out to be false, then we have evidence suggesting that the hypothesis is false—there is no ball.

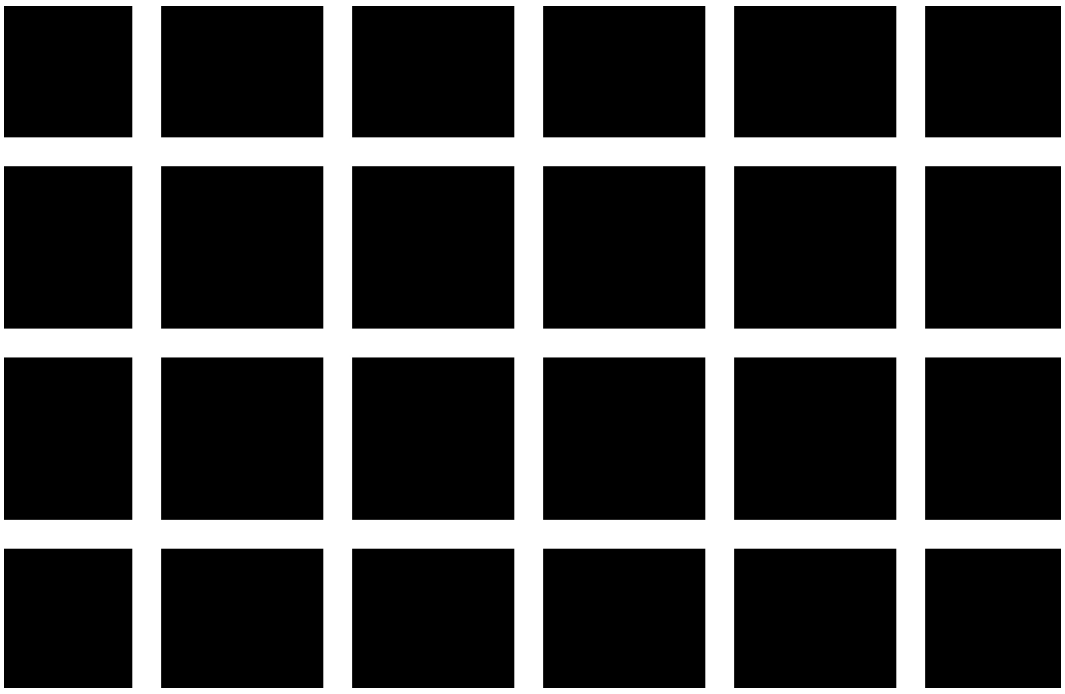


Figure 2.10: The Hermann grid illusion. The appearance of scintillating grey dots, at the intersections of the white lines, occurs only in our minds (it is produced somewhere between the retina and our conscious experience of the image) rather than in the outside world (somewhere on the page or between the page and the retina).

One way to notice that the earth is not flat is to lift oneself up from the surface of the earth and to look at it from a distance (see Figure 2.11). One way to notice that the sun does not go around the earth is to stand back from them both and watch how they behave in a richer context (including, for instance, the other planets). One way to notice that the color of the sky is not a constant blue is to pay attention to its changing tint as one moves one's eyes from the horizon to overhead. Another way is to watch it over the course of a day; during a sunset the blueness disappears entirely as the sky instead takes on shades of yellow and red.[77]

[77] And this effect isn't caused by redness obstructing our view of the naturally blue sky. The sky is red for the very same reason it is usually blue—more of those wavelengths are being filtered out of white sunlight and bent towards our eyes. It is possible for the blue sky to actually turn red because it was never *intrinsically* blue in the first place.

Figure 2.11. The earth, with a diameter of nearly eight thousand miles, viewed from a vantage point about twenty miles above its surface. The curvature is slight but obvious and, according to the photographer who took the picture, not an artifact of lens aberration. Note also: from this perspective the sky happens still to be blue.

In short, we can recognize an illusion any time we discover that our observation of a pattern has systematic exceptions. We can know that the pattern in our mind is not in the world when that pattern arises, at least in part, as a result of a limited perspective, or a limited perceptual system, or a limited mode of interaction with the world, and when the pattern fails to arise as that perspective, perceptual system, or mode of interaction changes in some way. And we can determine that one of these limitations has been afflicting us whenever we find we can reliably manipulate some aspect of

the perceptual situation—but not any aspect of the apparently observed pattern itself[78]—and thereby find a way to reliably cause the pattern to disappear.

As simple as that may sound, the point has been lost time and again. It was lost on those who, faced with Copernicus' evidence, still believed the sun to go around the earth. And it is lost on anyone today who might still believe objects to truly *have* constant colors or weights (or those who believe jokes to be inherently funny; see, *e.g.* the discussion in Hurley, Dennett, and Adams 2011). Illusions often confound us in spite of reason. We may willingly ignore the circumstances (such as a sunset) that make a pattern (such as the sky being blue) disappear, and by doing so we may let ourselves fall for an illusion. I'll spend most of Chapter IV and part of Chapter V trying to show that our common intuitions about the existence of proper functions have systematic exceptions that are just about as easy to recognize as sunsets, and yet which many of us are somehow unwilling to take as evidence that proper functions are a kind of illusion.

Evidential Onus

As I said, however, my work in Chapter VII will cut the other way. Rather than trying to prove that a widely observed pattern *is* an illusion; I will try to make the case that a widely observed pattern *is not* an illusion.

The important contrast that warrants emphasizing "is" and "is not" does not lie in the difference between proving and disproving (that is a critical but distinct issue in science and logic), but rather in the difference between going with and going against the weight of evidence. As Hume showed us long ago, when there is debate between parties about whether to believe some claim, we

[78] Of course one wouldn't want to manipulate the pattern itself—the point is to see if the pattern is really out there in the world; damaging or destroying the pattern *ensures* that it no longer is. To repeat, what we want to do is only to alter other parts of the world and our interaction with the world to see if the pattern disappears as a result of those other factors.

can often assess which side of the debate ought to have more work to do in order to make their case convincing. The answer is simply that the side whose claim is initially less credible (given the prior evidence) must provide proportionally more new evidence to support it (Hume 1748; see also Laplace 1812).

For example, we should expect it to take more than one mail-order catalog advertising dry-germinating seeds to convince an old farmer that there is a new and different way to grow crops during a drought. Even in drastic times, the farmer won't even blink before dismissing the advertisement out of hand. The salesperson or charlatan who is selling the dry-germinating seeds has an enormous burden of proof to overcome, and the onus is on them to powerfully prove that their incredible seeds work.

Hume's own example was the idea of life after death: Since we know with great certainty, from the testimony of a culture made of millions or billions of historical observers, that no person has ever come back from the dead, then it should take much more than two or three earnest-sounding-people's testimony to overthrow that conviction. It should take, Hume points out, even more than our own direct witnessing of an apparent corpse returning to life, since the likelihood that we have been deceived in the one miraculous-seeming instance is vastly greater than the likelihood that we and everyone else have been deceived in the myriad-myriad episodes that formed our prior conviction.[79] The onus is on those who would like us to believe in life after death, to provide, as Carl Sagan put it, *extraordinary* evidence for their extraordinary claim (Sagan 1979).

In the case of the illusion of color constancy, the prior evidence that things have colors was (and still is) pervasive, and so color constancy is in fact the reasonable thing to believe, at first. The onus therefore was in fact upon the scientists of the seventeenth, eighteenth, and nineteenth centuries to provide evidence of how the phenomena of color arise so regularly *without* being

[79] As the reader may suspect, this notion can be, and has been, cast in terms of Bayesian reasoning (see, *e.g.*, Borges and Stern 2007; Brown 1970; Madruga *et al.* 2003; Pigliucci and Boudry 2013).

intrinsic properties of objects (and they did this admirably). In the case of proper functions, the most common prior belief today is also that they do exist, and so if I'd like to show otherwise, then the evidential onus in fact rests on my shoulders. This is why, in Chapters IV and V, I must put dozens of pages of effort into producing real examples that (hopefully) can alter our perception and thereby dispel the illusion.

On the other hand, as I'll argue in Chapter VII, the pervasive prior evidence about the phenomenon of goal-directedness is also that it is a pattern that does indeed exist, and so, quite similarly, the onus should not be on me or any other believer to prove that it does exist. The burden of proof should instead rest on the shoulders of those who disbelieve in that pattern to provide overwhelming evidence to the rest of us that our common impression of the phenomenon is an illusion. The onus is on them to find a way to change our perspective on organisms—without changing the organisms themselves—and thereby to cause them to no longer appear goal-directed.

E. Value

Equilibrium systems fail to be genuinely goal-directed when their equilibrium maintaining behavior is of no value for anything.

—Mark Bedau (1992a, p. 38)[80]

In the previous chapter we started to look at the sometimes value-laden topic of normativity. There, I distinguished between, on the one hand, objective, comparative norms (such as whether Ceres ought to continue along its orbit around the sun tomorrow) and, on the other hand, subjective, evaluative norms (such as whether we ought to keep the fish we've caught cool until we intend to eat them). I pointed out that the evaluative type of norm is of central importance to the notion of goal-directedness. There can be no goal without some kind of accompanying subjective standard by which one could evaluate whether the agent whose goal it is has or has not achieved the goal. That is to say, evaluative norms can explain (whereas comparative ones cannot) what it is that ought or ought not happen in order that an agent can be said to have achieved its goal.

The notions of value and evaluation are intimately tied in with teleology. Achieving a goal is *good* for the goal-directed agent, and not achieving it is either neutral or, more often, *bad* for that agent. Part of understanding teleology in its entirety will mean understanding what it means for something to be good or bad for a goal-directed agent and its goals. Thus a theory of teleology needs to be paired with an explanation of how benefit may accrue to an agent and how that benefit may be evaluable.

[80] Bedau's use of the word "equilibrium" here refers not to thermodynamic equilibrium, but to state-maintaining (cybernetic) systems, a topic we'll look at in Chapter III.

At present, however, the notion of value is a deeply metaphysical one, over which there is plenty of debate amongst philosophers. It stands at the core of both ethics and economics, each of which has been studied extensively over the ages, and yet, despite these millennia of analyses, it is so far unclear just what physical features—or physically emergent properties made of space, time, atoms, and forces—could account for it. Just as is the case with goal-directedness, there is a prior, objective world—the world of physics and chemistry but not of agents and psychology—in which value simply does not exist, and then there is the full-fledged biological, psychological, economical world, which, in large part, revolves around the perception of, striving for, and trade in value. The question is: by what means did the former transform into the latter? From where in our world did the phenomenon of value emerge?

An Energetic Hypothesis about Value

From the definition of a dissipative system, one might come to think that value may be a function of energy or, at least, of usable free energy. Indeed, I happen to have run across a recent theory of economic value that makes just this presumption: "Organized forms of matter, such as organisms, are examples of dissipative structures that feed on the opportunity to create a flow of free value to maintain their structure and to grow and develop." (Roels 2012). What Roels means, when he describes freely flowing value, is the ability of dissipative structures to exploit the free energy from within an external energy gradient that flows through an environment (for instance, solar radiation, here on earth) to extract work (see also Costanza 2004, for a similar perspective).

I admit that it seems reasonable at first to think that if free energy is irreplaceable in its role in creating the valuable structures that make up organisms and their artifacts, then it just may be the fundamental currency of value. After all, one might also note that trade in energy is the most

significant cornerstone of global financial markets (especially if we consider not only industrial sources of energy such as oil, coal, and natural gas but also biological sources such as wheat, rice, and corn). However, as is so often the case, first impressions can be misleading.

To begin with, there is another freely available feature of the world that dissipative structures and processes feed on hungrily and depend on irreplaceably in order to maintain their structure and to grow and develop. I am speaking of course of atoms. Without a constant flow of available matter, dissipative systems could neither grow nor rebuild lost structure. Perhaps we should say instead that it is atoms that are the source of value?

Or perhaps we should fork Roels' theory, so that one tine of value might be constituted by free energy while another consists of matter? With a fork that accounts for the contributions of both material and activity, we could seemingly have our metaphysical cake and eat it too. But there is an even more fundamental problem with this account that the forking strategy leaves unresolved. Even if freely flowing atoms and energy were both elementary sources of value, that would mean that anything that came to be, through the interaction of atoms and energy, would be a beneficiary of value. The problem here is that that describes literally every pattern in the entire universe. There would be no distinction between the subjective and the objective—no account of what delineates a particular kind of pattern as being the *agential* kind, to which notions of value and benefit may apply, in contrast with the remainder of non-agential patterns, for which value judgments are simply inappropriate or irrelevant.

A Temporal Hypothesis about Value

In order to elucidate the distinction between the objective and the subjective (and also to account for the evaluative norms that underpin teleology), we are going to need a different

hypothesis about value and benefit. Value will need to depend in some way on a resource that is differentially available to subjective and objective patterns in the world—a resource that is important to the one and irrelevant to the other. The answer that I believe supports this distinction, and that I'll argue for in detail in Part II, is *time*.

In short, the idea I'll offer later is that free energy and material are not *intrinsically* valuable, but they can *become* valuable (and then can be evaluated) when used in the right way—namely, by living, teleologically organized patterns—in order to buy more time. I'll argue that what teleological patterns do is to use both free energy and material in order to perform constructive work . . . but not just *any* constructive work. They use these resources to perform the specific constructive work that ensures those same patterns—that is, themselves—continue to exist. In this regard, teleological patterns can be put in sharp contrast with both spontaneously organizing patterns and ontological nonce, both of which may also consume free energy and material in undergoing structural changes, but neither of which is able to buy itself more time. The difference can be found in the nature of these categories. Spontaneously organizing patterns have all the time they need from the get-go—they are bound to exist as long as their energetic and material precursors (co-) exist. Stars and planets and crystals simply form when the conditions are right. In contrast, ontological nonce, by its very definition, has no time—such patterns are bound not to exist more than momentarily,[81] except as fleeting intermediate states in the constant ebb and flow of universal change. The notion of value applies neither to patterns that are *bound* to exist nor to those that are bound *not* to exist; it applies only to patterns whose existential fate can change. As we will discover, that is just what teleological patterns are.

[81] This is a broad sense of the term "momentarily" that can only be understood relative to the local background distribution of braising energy.

F. Identity

The ship wherein Theseus and the youth of Athens returned had thirty oars, and was preserved by the Athenians down even to the time of Demetrius Phalereus, for they took away the old planks as they decayed, putting in new and stronger timber in their place [so completely] that this ship became a standing example among the philosophers, for the logical question of things that grow; one side holding that the ship remained the same, and the other contending that it was not the same.

—Plutarch, *Theseus* (75 AD)

The question of identity is one of the oldest problems in metaphysics, and the paradox of Theseus' ship, as described above by Plutarch, is one of the clearest illustrations of the problem. Just what physical criteria can we use to determine the identity of the ship, or of any other object in the world? If atoms (or planks) are constantly being shed from and adjoined to objects[82], then what possibility could there ever be of giving a definition of an object in enduring objective terms? How could something ever be precisely the same over time?

In this work, I am going to propose what I believe to be a complete answer to the ancient problem of identity. There can be no objective theory of goals or of goal-directed agents without an objective theory of what constitutes the *agent* whose goals they are, or who benefits from the achievement of those goals. The theory of teleology that we will examine in Part II is one in which certain types of patterns are, thanks to a special trick in their organization, able to maintain their own identity in the face of constant braising.

[82] And, at some time scale, this is certainly the case for every object in existence. Fluids flow, liquids evaporate, gases condense, metals corrode, amorphous solids sag under gravity . . . Even more ubiquitously, heat and background radiation are constantly braising objects, causing sometimes tiny but nonetheless generally irreversible rearrangements of atoms.

My answer to the general question of identity, however, comes in two independent halves, which we can distinguish by using the labels *identity* and *identifiability*. The idea is that some patterns in the world have true, *objective* identities, while the rest are merely *subjectively* identifiable, relative to some observer's purposes. Identity and identifiability are both ways to group various patterns together according to some measure of sameness, but I'll argue that they do so on fundamentally different grounds, which can be distinguished primarily in terms of *purpose*. I'll try to clarify that distinction by looking more closely at each one now.

Identifiability

The notion of identifiability relates to the apparent identity that an act of categorization gives to an object such as Theseus' ship or, for that matter, most random lumps of material floating anywhere in the universe. Identifiable things do of course have a kind of sameness to them; that's what makes them identifiable. But that sameness has no independent objective measure. It is a sameness that depends upon the identifying intentions—the purposes—of a teleological observer.

The ancient Greek philosopher Heraclitus famously emphasized the absence of *identity* in things that are *identifiable* with his quip "no man steps into the same river twice". While it is true that rivers are always changing, it seems philosophically troubling to claim that the Delaware that we see today is not the same river that General Washington famously crossed with his troops during the American Revolution, or that the Nile that the historian Herodotus wrote about is not the same river that the ruins of the ancient capital of Memphis still stand beside. Despite the changes in the courses of these rivers, we feel that they are the same rivers they once were. That is, we can still *identify* these and other rivers, as long as we assume some subjective criteria, the contravention or satisfaction of which can be used to judge whether any changes in a particular river are of a

sufficient scale or kind to warrant either dismissing or retaining our judgment of its (relative) identity. If a dam diverts a river and the water that would have arrived in one place now takes a somewhat different path, we are free to decide whether or not, according to our criteria—that is, *for our purposes*—it should be called the same river.[83]

In other words, *we* (and other agents) are able to identify objects based on our choices of *subjectively* essential features, but *they*—the objects—don't have any kind of *objectively* essential features that can be universally relied upon for identification over time. Nothing in particular—no solitary essential feature or group of features—needs to remain the same about such objects in order for them to have a relative identity; they simply need to remain similar enough to continue serving some agent's purposes, so that that agent will continue to be willing to identify the object as the same thing (Wittgenstein 1953).

Although things that have true identities are also typically identifiable, the majority of identifiable things can be equated with the earlier-described category of ontological nonce. That is, they are things that may have an informational description at any one moment, but they are subject to braising, and thus ultimately can only be maintained to be the same—according to any subjective measure by which they are claimed to be the same—through the intentional efforts of an agent who desires that maintenance. Absent any such efforts, they are bound to exist as one pattern at one moment, and a different pattern at the next. Like Theseus' ship, which goes through a continuous series of damages and repairs, identifiable patterns will be transitory, ephemeral things.

[83] This phrase, "for our purposes" (or, similarly, "for current purposes" or "for the sake of this discussion"), allows us to create an ad-hoc category, often labeled with a word—old or new—that serves to bring a group into agreement as to what the criteria are that will be used to identify a particular thing. The rhetorical tool is necessary only if the criteria are not already publicly agreed upon—for instance, in the dictionary or encyclopedic definitions of a term. After all, cultural conventions such as words serve the very same purpose, just with a longer and broader established history of agreement.

The Hard Problem of Identity

The solution to that half of the problem of identity is relatively simple, and the reader will have noticed that the concept of *purpose* lies at its heart—identifiability exists only with respect to a (real or imagined) purposeful identifying *agent*. But this leaves unresolved the question of what accounts, in the first place, for what I called *identity,* which brings into existence those subjective, teleological agents who have the capacity to identify things for their purposes. It cannot be that the identity of such agents is also relative to some other subjective external judge, or else the logic would be regressive, with no answer ultimately given.

We can call this second, deeper quandary "the true problem of identity" or, mimicking David Chalmers' (1995, 1996) distinction regarding various problems of consciousness, "the *hard* problem of identity"[84]. An answer to the hard problem of identity stands at the hard core of the modern theory of teleology that I'll describe in Part II. Although the short version of that answer, which I'll give now, may at first seem simplistic and tautological, that impression will fade when we begin to understand what it takes to realize that claim. The simple answer is that the identity of a teleological agent must, in some regard, remain precisely the same across time. In order to have an intrinsic identity of one's own, one must be *truly and objectively identical* to oneself. Something really must not change. This seems obvious, but it also seems inconsistent with the statistical, thermodynamic results of braising, whereby the atomic structures of objects are constantly being reorganized. However, in Chapter IX, we will look at a more abstract formulation, given in terms of both *organizational information* and *time*; that formulation will be able to make rigorous the notion of an item's remaining truly identical, despite relentless environmental braising.

[84] The hard problem of identity deserves the name "hard" because it has gone unanswered for millennia; however, I believe it to be solvable in materialist terms (and I will make an attempt to solve it in Part II of this work). This can be put in contrast with Chalmers' own "hard problem of consciousness", which he has dubiously labeled "hard" because he claims that there is no materialist solution to the problem and that we have no recourse but to resort to an immaterial (dualist) solution.

Revisiting the Ancient Paradox

We can return now to the problem of Theseus' ship and the "easy problem" of relative identity. At the cost of requiring a theory of teleology in our world (which is a price I am prepared to suggest we can pay), our theory of identifiability resolves the ancient paradox by rejecting the question outright: The ship *never had* an individual (true) identity in the first place, and so, when it changes in its material constitution, it *still* does not have such an identity. Instead, the ship, like any object, always had the possibility of fulfilling any number of identity roles with respect to various subjective, teleological agents' purposes in making the identification.

Crucially, if two (or more) agents have differing purposes for considering the ship to be a ship—if it serves different functional roles in the different agents' conceptions—then when the ship changes materially in any way, its identity from each perspective will have to be re-evaluated, resulting potentially in conflicting opinions. After most of the ship's planks have been replaced, Theseus himself will likely consider the new bundle of freshly installed planks to be his own ship, but ontological pirates—crazy imaginary thieves who would attempt to steal by merely reinterpreting the identities of items in deeds and titles of ownership—may consider this floatable bundle of planks to be a new ship, up for grabs, and they may even consider Theseus' ship to lie in a corner of the shipyard in the form of an abandoned pile of old disassembled planks. Both claims are subjective claims of identifiability that are made relative only to one's purposes; *there is no objective truth to the matter*. The reason Theseus can win this argument is not that his claim is more objective, but rather that he has also paid for the purchase and installation of all the new planks, and so those new planks,

in whatever form they take, are actually his, too (and he has the rule of law on his side, so the society will back him up).[85]

The Fundaments of Subjectivity

We've spoken now about both *value* and *identity*, two concepts that—much like the topic of goal-directedness itself—absolutely permeate the subjective realms of biology, psychology, economics, and so on, and yet never show up in the objective worlds of pure physics or chemistry. The phenomena behind these two presently metaphysical concepts form what I will call *the fundaments of subjectivity*. They both are involved in every facet of our world that can be labeled teleological or agentive; and teleology or agency is involved, in one manner or another (typically as the subject), in every phenomenon that can be labeled as subjective.

To make it clear, my claim is that a teleological pattern can only exist if (i) it has an identity, meaning that its goals can be said to be its own, and if (ii) contributions can, in principle, be made toward the satisfaction of its goals—if it can benefit by its goals being achieved. Without these fundaments, there can be no *goals,* and thus no *goal-directed patterns*. Without these fundaments, there can be no agents and no agency—no actors that can strive to attain their goals. Without these fundaments, the subjective aspects of the world do not exist.

My contention is that solutions to the metaphysical problems of value and identity are key factors in finding a theory of teleology. These are the main problems that challenge us as we seek to bring subjectivity, teleology, and agency under scientific consideration. And so, on my view, any

[85] The sorites paradox—named after the Greek word for "heap"—asks a question similar to that of Theseus' ship. We might notice that the removal of any single grain of sand from a heap leaves behind "the same heap", and yet, if the process is repeated enough times, the heap will complete disappear. The theory of identity I offer resolves this paradox, too, by claiming that the heap of sand never had an objective identity to be lost—it was always *merely identifiable*, as a heap relative to our purposes or intentions. And so, just as we used our judgment to define it as a heap in the first place, we are also free to draw a subjective line that defines when it no longer is a heap.

teleological theory ought to focus primarily on understanding the kinds of material organization that account for the phenomena underlying these two concepts.

G. Agency and Natural Selection

The replicator approach [to understanding natural selection], in many versions, is designed to mesh with an "agential" way of looking at evolution, a perspective in which we see the entities in an evolutionary process as pursuing goals, having interests, and using strategies.

—Peter Godfrey-Smith (2009, p. 36)

To be clear, in the quote above, Godfrey-Smith is elucidating a viewpoint that he and a large number of eminent evolutionary biologists do not endorse. Godfrey-Smith does appreciate the modern view that *genes* seem to be a more likely unit for natural selection to act upon (at least more so than *species*; I'll say more about this debate below), but he resists viewing genes as *agents* that would, as he says, have goals, interests, and strategies. A number of our later discussions will refer to natural selection, both in and of itself and in its relation to agency, and so I need to introduce those concepts now[86]. Although I am placing natural selection in this chapter alongside thermodynamics, existence, causation, identity, value, and so on, I want to avoid giving the impression that the process of evolution is in some way as fundamental to teleology as are those other topics. I think an impression of that sort would be an illusion. There is, of course, intuitively, a very deep relationship between evolution and teleology (since both have profound connections to life and biology), but I think the usual perception of that relationship is entirely upside-down.

Today, the popular way to understand the link between evolution and teleology is that the purposeful things that exist in the world are the products (or, in the case of artifacts, the products of

[86] The process of evolution itself is also sometimes imagined to be agentive, as the development of traits that serve purposes seems to require some intention for those purposes to be served. Today, this is widely understood to be a false impression. I will address and dismiss the notion in more detail in the next chapter.

the products) of natural selection. Under that conception, evolution is a natural (non-human) designer whose work, over the ages, comes to create all the reasons and purposes in the natural world. The eye was designed for sight; the wing was designed for flight. A thing is *for* something—it serves a purpose—if it has been *designed* for it (see, *e.g.,* Allen and Bekoff 1995a; Dennett 2014; Godfrey-Smith 1993, 1994; Griffiths 1993; Kitcher 1993; Millikan 1984, 1989a, 2002).

I think this viewpoint is so enticing that I will spend quite a number of pages in Chapter IV attempting to debunk the intuition it is based on, which I call "the design fallacy". Certainly there is a sense in which most biological traits have been crafted in part by natural selection, and so whatever purpose these traits serve owes its existence in part to natural selection. However, when I expose the design fallacy, I will conclude that rather than viewing evolution as giving rise to teleological patterns, we should see things in precisely the opposite way: being teleological is one of the rare ways[87] that a pattern can stick around in this world long enough to be subjected to natural selection[88].

A pioneering version of the agency-based view, initiated by William Hamilton (1963, 1964) and George Williams (1966), though developed most completely by Richard Dawkins (1976, 1978, 1982), claims that replicators—the underlying substrate upon which the process of natural selection operates—are themselves agentive, teleological, and natural potential beneficiaries. Dawkins' (1976) title, *The Selfish Gene*, highlights the idea clearly: germ-line genes (not organisms, and not somatic genes) are the ultimate replicators in biology . . . and their behavior is selfish. That is to say, these genes are agents (they have selves) whose many extended behaviors are performed ultimately for their own benefit; whatever it is they do is done *in order that they may replicate*. I want us to take note

[87] As I see it, the other potential way of sticking around long enough is by being a spontaneously organizing system, but even then, only some such systems have what it takes to be subject to natural selection.

[88] One might protest, "Couldn't both be true?", and I think it a fair question. But once we see the theory of teleology, in Part II, we'll have more reason to see that a teleological nature of things can in principle exist independently of the selective processes that later come to work upon those things.

of the *normative* aspect of these assertions, and of the "in order that" clause emphasized above; this is an absolutely teleological claim: the *goal* of replicators is to replicate.

The assumption underlying this view is that genes are the things that are most faithfully and continuously reproduced in biological reproduction, and so it is only they that can be said to benefit—in terms of longevity—from successful replication, and it is only on them that natural selection must ultimately be operating. On the contrary, suggests Dawkins, organisms, as the sometimes extended and complex phenotypes of those genes, should be seen not as *replicators* but as temporary, replaceable *vehicles* that are driven about in the world by the actual replicators. Such vehicles interact with the environment *on the replicators' behalf* (Dawkins 1976, 1978, 1982; see also Dennett 1995; Haig 1997; Hull 1978; Maynard Smith 1998). He writes:

> Four thousand million years on, what was to be the fate of the ancient replicators? They did not die out, for they are past masters of the survival arts. But do not look for them floating loose in the sea; they gave up that cavalier freedom long ago. Now they swarm in huge colonies, safe inside gigantic lumbering robots, sealed off from the outside world, communicating with it by tortuous indirect routes, manipulating it by remote control.
>
> They are in you and in me; they created us, body and mind; and their preservation is the ultimate rationale for our existence. They have come a long way, those replicators. Now they go by the name of genes, and we are their survival machines. (Dawkins 1976, pp. 19–20)

The story here is gripping. These little creatures—genes—who once replicated on their own (perhaps in an RNA-world scenario[89]) are the agentive, teleological, beneficiaries of their own behaviors. And having become highly developed through evolutionary processes, their main activity today, in order to promote their own goals (of replication), is to build and then direct machines that do most of the dirty work of survival for them. They are *managers*, *par excellence*.

I don't want to relinquish the powerful intuition in this picture. However, this very intuitive view (now called the "gene's-eye perspective") has come under reasonable, fervent attack from other scientists (like Godfrey-Smith, above), who also find it powerfully intuitive that simple material objects such as genes simply can't be the kind of agents Dawkins would want them to be—they don't have goals or subjective perspectives, and they aren't subject to an evaluative kind of normativity. There is nothing these molecules *ought* to do, and there is no way that anything might be objectively *good* or *bad* for them—no more than for any other molecules. Genes are just certain nucleotide polymers that undergo interesting physical and chemical interactions when located within the context of intercellular machinery, but that are otherwise relatively inert. That is, they might play a role *within* agents, but they can't themselves *be* agents (see also, *e.g.*, Brown 1998; Symons 1981).

Considerations of this sort have led Dawkins and a cohort of sympathetic theorists to moderate their claims, suggesting that genes aren't literally agents, but that the most useful way to think about how evolution works is to treat them as agents (Dawkins 1976; Dennett 1987, 1995; Haig 1997; Maynard Smith 1998). This scaled-back version of the agentive claim represents a conciliatory middle ground just where one is needed. One must reckon both with the perception of agentive activity and with the materialist reality at the same time. But in taking all the bite out of Dawkins' powerful intuition, I think a compromise of this sort gives too much to the other side.

[89] This origin-of-life scenario considers that RNA can serve not only as a replicating molecule but also as an enzyme for the replication of other RNA strands. In cooperation, groups of various enzymatic RNA strands could help one another replicate without needing the other machinery that makes up cells (see, *e.g.*, Crick 1968; Gilbert 1986; Kauffman 2000; Maynard Smith and Szathmáry 1995; Orgel 1968; Woese 1967).

Dennett (*e.g.*, 2005) offers a position that is a step closer to what I will suggest. For Dennett, it is important to note that genes aren't molecules of DNA. They are strands of *information* that are usually *embodied* as molecules of DNA; it is the information, not the physical instantiation, that is central to their nature. While I appreciate the abstraction in this perspective, I think it still doesn't go far enough. Genes defined informationally still don't have agentive qualities on their own; they are still inert except in the context of machinery that can operate upon them, and we must still take them as agents only metaphorically.

We can locate a different middle ground by agreeing with Dawkins and Dennett that there should be *some* pattern that is minimally agentive and teleological and whose behaviors can be said to be for its own benefit, while also agreeing with critics of the gene's-eye perspective that such a minimally teleological (or agentive) pattern cannot be so simple as a strand of DNA or even an informational gene.

What we're going to see in Part II is that the teleological patterns I will describe there can fill this role. They can be the literal agents that are required to make Dawkins' account of lumbering robotic survival machines work. Those patterns are able to benefit, and their status as agents comes from the fact that they have true identities and can behave on their own behalf, thus providing their own benefit. This notion can replace that of a gene as a replicator[90] while still supporting the literal version of Dawkins' normative and agentive claims, thus leaving the rest of his marvelous picture intact. At the same time, it shouldn't offend the materialist sensibilities of his critics, because the account won't rely on an indefensible notion of agency within something so simple as individual molecules or strands of information.

In the meantime, now, I'll further outline the classical statement of natural selection, and some relevant modern adjustments to it.

[90] Or, really, as a *persistor*, replication being seen as a principal variety of persistence.

The Rudiments of Selection

The basic process of natural selection is widely understood today but, to avoid any possible misunderstandings, I will briefly sketch it out here. The classical definition is that evolution by natural selection is the process by which reproducing populations change in their statistical makeup whenever (i) there is variation among the members of the population; (ii) that variation can be passed from parents to offspring; and (iii) different variants have differing levels of "fitness" (as measured by the number of offspring, or sometimes grand-offspring, that they can be expected to produce). The process ensures that the variants that have higher reproductive success rates will (in relative terms) flourish, as compared with those that have lower rates, which will (again, relatively) flounder. *Selection* itself is the process by which variants live or die according to their fitness, resulting in a degree of representation within the population that reflects that fitness. *Evolution*—paradigmatic evolution by natural selection—is said to occur when this process of selection is combined with a moderate rate of mutation or other source of variation, and when that variation is able to create individuals with incrementally greater fitness, the overall result being turnover in the population, thanks to which new traits and, ultimately, new species may come to develop.

Essentialism

The above may sound fairly straightforward, but things are never quite so simple. Among scientists, there is no dispute that the classical statement is what ultimately accounts for the development of complex traits such as eyes, brains, and circulatory systems. However it is also becoming widely recognized that the classical statement is an *essentialist* claim, and that there is good reason to believe that it doesn't reflect all instances of population behavior. A more modern

statement of natural selection was given in recent years by Peter Godfrey-Smith (2009), who has attempted to amend the classical definition by recasting it in terms of *Darwinian populations*—groups of interacting individuals that have graded attribute membership along a number of dimensions[91]. On Godfrey-Smith's account, some such populations are central, paradigmatic cases wherein all the factors that lead to full-throttled evolution are aligned, while other populations are more minimal or limited cases, differing on one or another dimension, leading often (but not always) to some degree of selection. Furthermore, depending on various features of the population, even when there is selection, there may not always be evolution.

Later, when we review theories of function and teleology that rely on natural selection in their specification, it will be useful to recollect the various ingredients that make up selection (reproduction, variation, heritability, and differential fitness of variants) and to ask ourselves which of those ingredients a theorist might be claiming underlie teleological phenomena, and whether the theorist has some reasoned analysis underlying such a judgment. Can individuals that are unable to reproduce have functional parts or behaviors? If a developmental mutation turns out to be useful to an organism but goes uninherited, does it serve a purpose for that individual? In general, if one claims that natural selection grants functions to traits, is it the whole recipe that matters . . . or should our theories focus only on some subset (and, if so, which subset?) of the ingredients that make up Darwinian populations? In the end, I'll claim that the entire endeavor is misguided; however, in making that difficult diagnosis clearly, I will have to hold some theorists' feet to the fire by asking just what it is about natural selection that they think has teleological power, and why.

[91] The particulars of these dimensions are interesting, but not entirely relevant to my work here, which does not center on evolution but only hopes to understand the relationships between various selection processes and the teleological patterns that often take part in them. However, for those who are curious, the first dimensions that Godfrey-Smith analyzes include the fidelity of heredity, the abundance of variation, the competitive interaction with respect to reproduction, the continuity (or smoothness) of the fitness landscape, and the level to which fitness differences depend upon intrinsic properties of the individuals of a population (2011, p. 63). Variations along each of these dimensions produce decisively different kinds of Darwinian populations, which thus evolve in differing manners and to differing degrees.

H. Proto-Physics

If we take in our hand any volume of divinity or school metaphysics, for instance; let us ask, Does it contain any abstract reasoning concerning quantity or number? No. Does it contain any experimental reasoning concerning matter of fact and existence? No. Commit it then to the flames: for it can contain nothing but sophistry and illusion.

—David Hume (1748)

If the aim of physical theories is to explain experimental laws, theoretical physics is not an autonomous science; it is subordinate to metaphysics.

—Pierre Duhem (1906)

In the epigraph at the start of the chapter, I quoted an entry from the *Oxford English Dictionary* that defines metaphysics as the study of many of the topics that we've now begun to explore. The OED's next entry amends the first by claiming metaphysics is "the study of phenomena *beyond the scope of scientific inquiry*." I don't agree with the connotation of permanence in the phrase "beyond the scope" but nonetheless I prefer this latter definition because of its generality.

The definition that lists various member topics of the field is problematic because that list of topics has been prone to undergoing constant revision throughout the history of philosophy. Over the millennia, phenomena that were once seen to be topics of metaphysics came to eventually be understood scientifically, while other phenomena once thought to be topics of physics came to be reclassified as metaphysics, based on judgments that those phenomena were unmeasurable or

unfalsifiable (van Inwagen and Sullivan 2016).[92] In this regard I see metaphysics as a placeholder—a kind of *ad hoc* category into which we tend to toss the study of any phenomena for which we don't yet have an empirical approach. However, it might be fruitful instead to categorize these topics as *proto-physics* (or else to consider the term "metaphysics" to simply *mean* proto-physics) in order to emphasize some hope for our future potential to understand them.

Hume's empirical stance in the epigraph above is of course well taken; we should certainly put stock largely in ideas that we can numerically quantify and experimentally verify as being (approximately) correct. But the value in metaphysical (or proto-physical) practice is not in its always being able to offer correct theories. It is rather in giving us a space wherein, and a set of discursive tools with which, we can cast our nets broadly in search of correct theories. It is then up to science to figure out how to take those exploratory ideas, make them testable, and figure out which ones are closest to correct.

So the fact that these topics are not today subject to empirical scientific study doesn't mean they are unscientific; it only means that these are phenomena for which empirical science is still in need of guidance, from philosophy and theoretical science, in discovering how to approach them. One day, I suspect, most of the topics now considered to be parts of metaphysics—especially those topics closely linked with subjectivity and teleology, and addressed here in this chapter—will be topics of study in physics itself, or in science more generally.

[92] I regard judgments of unfalsifiability to be just that: judgments, which are themselves subject to fallibility; and so I generally regard reasoning that uses such judgments to conclude that some particular idea may be unscientific to be simply unimaginative. It is not clear to me that any topic for which there is a pattern to observe and explore could be *fundamentally* unscientific—only, perhaps, *technically* and thus *temporarily* unscientific.

Chapter III

Finalism and Vitalism: A Brief History of Western Teleology

The notion of teleology arose most probably as a result of man's reflection on the circumstances connected with his own voluntary actions. The anticipated outcome of his actions can be envisaged by man as the goal or purpose towards which he directs his activity. Human actions can be said to be purposeful when they are intentionally directed towards the obtention of a goal.

—Francisco Ayala (1970:8)

Teleological thinking has had a checkered existence, cycling through phase after phase of acceptance and rejection throughout the history of science and philosophy,

(+) from ancient animistic teleology

(–) to the mechanical purposelessness of early Greek materialism and atomism

(+) to both Platonic and Aristotelian finality[93] and, later, the Christian theological teleology that lasted through the middle ages

(–) to the teleological dismissal rooted in the mechanistic ideas of the scientific revolution

(+) to eighteenth and nineteenth-century vitalist views, and then

(–) to the eventual abandonment of vitalism.

In this chapter, I'll review this history of teleological thinking up until about the nineteen-thirties. The history can be divided into two parts, both of which span the same period yet have often been

[93] An early term for goal-directedness that, as we'll see shortly, puts purposiveness in terms of causality rooted in *ends*.

treated as separate topics: *Finalism* is broadly concerned with goal-directed phenomena in the world while *vitalism* claims to be more narrowly concerned with life and living processes (although, as we'll see, vitalist hypotheses are also steeped with goal-directedness). We'll first review the development of finalism, beginning from its pre-philosophical roots and continuing through until the scientific revolution. Then we'll return to review vitalism, tracing that tradition also from its ancient metaphysical roots through to its interactions with the scientific developments of the past two centuries.

By the time vitalism was banished entirely, many expected that teleological thinking might finally disappear along with it. It seems that many still hope it will, or think it already has (see Chapter VII). But the cycles of acceptance and rejection didn't end there.

In the nineteen-forties, a new theory of goal-directedness—the cybernetic theory—was developed and it remained a well-regarded hypothesis for a number of decades before falling out of fashion later in the twentieth century. The basis of the cybernetic theory of goals is that any cybernetic system—one that involves a negative feedback process that homes in on a state—is thereby goal-directed toward that state (Rosenblueth, Wiener, and Bigelow 1943; Rosenblueth and Wiener 1950; see also Braithwaite 1953; Maxwell 1868; Nagel 1977; Scheffler 1959; Taylor 1950a, 1950b). The most classic illustrations of this are the stabilizing or homing mechanisms of a thermostat or of a heat-seeking missile[94], although other stock examples include the various kinds of homeostatic mechanisms in biological organisms. The cybernetic proposal satisfied materialist and mechanistic sensibilities handily because its explanation of goal-directedness was both plainly non-enchanted and clearly present in both artifacts and organisms; however, when subjected to analysis it

[94] The so-called Watt governor for controlling the speed of a steam engine is another commonly used instructional example.

eventually could not withstand the force of counterexamples, and at this point the hypothesis has been more or less abandoned.[95]

In the sixties and seventies, focus on teleological phenomena shifted sharply. As interest in the cybernetic theory waned, a literature debating the notion of *function* waxed, quickly taking hold of philosophical and scientific attention in biology. As we'll see, the majority of writers in this new (and still ongoing) tradition have, with few exceptions, turned out to be largely dismissive of goal-directedness. These topics from the latter twentieth century will form the central subjects of chapters IV, V, and VII. In the meantime, though, let's finish making our way through the previous two and a half millennia of teleological history.

[95] For instance, the behavior of a pendulum bob or a marble in a bowl, as analyzed in Chapter I, are both cases of feedback causing an item to home in on an end state, but neither is teleological.

A. Animism

Animism is by many regarded as the earliest form which religion took, and as the root from which was derived all religious beliefs which the world has known, and was also the earliest basis of all that is dignified by the name of culture.

—George William Gilmore (1919)

Human teleological thinking really begins with our own intuitions about things in the world that seem to behave as if they had a will of their own. At base we view ourselves, as well as our comrades and our enemies, as having goals and as attempting to see them through. But we also project goal-directedness out onto numerous other aspects of the world, some of which appear to be striving for something, and others of which we see as thwarting our own strivings.

It is easy to come to believe that capricious, unseen agents may be the causes behind some of the phenomena we don't understand. For instance, the motions and behaviors of dust devils and of storms, or of earthquakes, volcanoes, and tsunamis, frequently tempt us to see nature itself as being goal-directed. And the spooky feeling we get when doors seem to close on their own or when we hear rhythms that are lifelike (such as footsteps, when we *know* there couldn't be anyone up in the attic) is due to the mental projection of goal-directedness or agency—of spirits whose intentional behaviors we believe we may be witnessing—out onto the world. In cases such as these, when we are unaware of the invisible drafts, distant earthquakes, or sun-warmed patches of the sea that can genuinely explain the phenomena we observe, the notions of ghosts and gods can easily fill the explanatory gap.

We are even tempted to project goal-directed intentions in more ordinary circumstances.

When we are walking in a crowd and another person happens to step on our heel twice, successively, we sometimes jump to the conclusion that these paired accidents were actually done "on purpose". We might assume malicious goal-directedness and it is especially easy to assume because we know that people are generally goal-directed creatures, even if the double heel-step was only coincidental. At other times, there are persistent physical phenomena that occasionally appear goal-directed to us. For instance, when objects refuse to cooperate with us, such as statically charged hairs standing up despite our combing efforts or when a nut or bolt resists loosening, we sometimes plead with the stubborn objects as if their obstinacy were mischievously intentional.

Outside of the context of scientific inquiry and logical reasoning (both recent and rare modes of thought[96]), we humans tend to be overconfident in our ability to explain the world around us, often falling back simply on the most accessible explanations for the phenomena we observe (Rozenblit and Keil 2002). That explanatory confidence, in combination with the analogy from our own goal-directed behavior, makes it easy to project willfulness and desires upon various objects and parts of nature, especially in situations when no competing explanation—say, from a culture of science—is obvious. Spirits, ghosts, and gods, and other spirited entities emerge, from our imaginations, to inhabit the world.

Psychologists and anthropologists use the word "animism" to describe beliefs of this sort, which, they have found, show up commonly both in individual and cultural adolescence[97]. At the

[96] The scientific method was introduced to our culture only a handful of centuries ago (Bacon 1620) and formal logic, while being introduced at first a few thousand years ago, was only fully formalized during the past two hundred years (*e.g.* Russell 1918).

[97] It is reasonable to wonder whether animism also appears in alternative phylogenetic branches. Take apes, for instance—some of whom can limitedly use human language and who rudimentarily understand the sequential nature of numbers—do they project goal-directedness onto others? At the least, we might wonder whether they believe that patterns such as rolling stones or dust devils are alive. Perhaps a good set of experiments might one day discern how they behave toward simple, seemingly goal-directed robots. In the meantime we can speculate from the theory-of-mind literature since, to the extent that chimpanzees can attribute intentions and other mental states to one another and to humans, we can wonder whether they also can or do attribute those states to nonliving objects. But the literature on this topic is divided and the best evidence suggests that even if chimps or bonobos do have a theory of mind, the degree to which they have it pales in comparison to the highly productive tool that we humans have (Cheney and Seyfarth 1990; Call and Tomasello 2008; Hare *et a*l. 2001; Povinelli *et al.* 1990; Povinelli and Vonk 2003; Premack and Woodruff 1978).

individual level, the attribution of goals begins quite early. By six months of age, infants show an ability to distinguish between random motion and "biological" motion (Bertenthal 1993; Rochat *et al.* 1997) and somewhere between then and eighteen months they begin regularly inferring the goal-directed actions of people (Meltzoff 1995; see also Woodward 1998) and even attributing goal-directedness to animated shapes (Csibra 1998; Csibra *et al.* 1999; Gergely and Csibra 2003; Gergely *et al.* 1995; Keil 1994, 1995). Later, in what Piaget called the "pre-operational" stage of childhood development (roughly ages two to seven), children commonly make use of an animistic explanatory strategy in which they often assume that objects are much like people, and may characterize them as having feelings, desires, goals, and intentional behaviors (Piaget 1929, 1951; Piaget and Cook 1952). As we have all seen, even adults who "know better" frequently animate objects of many kinds—power tools and other machines are, in the minds of their users, imbued with the desire to cut us or crush us and so we treat them with respect so as not to be subjected to their ire; favorite cars or other implements are named and imbued with personalities; and we sometimes have lengthy monologues (some might say dialogues) not only with pets but even with plush toys. It is debatable how seriously these kinds of attributions are taken, but given that a great many adults still believe in gods and ghosts, it is not unreasonable to treat these other animistic beliefs with some degree of seriousness.

Research has found that eye gaze appears to be a central signal that helps apes make attributions of knowledge, attention, or intention, and so, at least at first blush, it seems unlikely that these animals would be able to attribute willfulness to faceless patterns such as stones or storms.

Figure 3.1: The volleyball named Wilson, who served as Tom Hanks' character's companion, in the movie *Cast Away*.

At the cultural level, anthropologists consider animism to be "an idea of pervading life and will in nature . . . a belief in personal souls animating even what we call inanimate bodies" (Tylor 1871). And cultural memory supplements anthropological research in reminding us that animistic beliefs of one sort or another existed in most early societies. We have all heard of tree spirits, water sprites, and other fairies; various angels and demons; the spirits that cause diseases; the spirit of the mountain or of the volcano to whom a child must sometimes be sacrificed[98,99]; or the soul of the forest, or the will of the sea, or of the storm; and so on . . . If we combine individual animism with a scientific naïvety and a cultural tradition of storytelling, one can easily see how a rich cultural mythos, and eventually a pantheon, might develop to explain the phenomena of the world in terms

[98] See, *e.g.*, Wade (2013) and Wilson *et al.* (2013), for various discussions of human sacrifice—*capacocha*—made to the mountain gods in Incan culture; Romey (2018) for exposition of a recently discovered Chimú site where Gabriel Prieto and John Verano have excavated the remains of over 140 sacrificed children; and also Gibbons (2012) for a survey of other sacrifices in cultures around the world.

[99] Even today, the legend of Pele the goddess of fire and of the volcano is still widely recounted and respected in Hawai'i.

of the spirits that direct it.

Although animistic beliefs play no role in the philosophical and scientific discourse about nature today (aside from as a *subject* in anthropological study), such beliefs are historically significant as they provide the backdrop for the remainder of the teleological thinking that has occurred over the millennia. They also draw our attention to the instinctually teleological lens through which we humans tend to perceive our world.

B. Finalism

Nothing happens in vain, but everything from reason and by necessity.

– Leucippus (as cited in Taylor 1999)

Pre-Socratic Greek Thought

As the earliest philosophers began to reason about the structure of the world, the first of these to have a significant impact on teleological thinking were the early Greek materialists and atomists. As we'll see, the account they arrived at, over the course of a few generations, has a surprising amount in common with modern materialism in terms both of its underlying physics and its anti-teleological leanings. In order to see how teleology is excluded from their account, let's quickly trace the development of these ideas from Eleatic materialism to Democritean atomism.

The Eleatic philosophers, in particular Parmenides and Melissus, first proposed their materialist account of the world as a reaction against the animistic and hylozoist views of their forebears. These philosophers proposed that the minuscule substances of our world are *solid*, *eternal*, *impassable*, and *indestructible* and, importantly, not living, willful, conscious, or soul bearing[100]. One peculiarity of Parmenides' view was that, by way of a rather contorted argument that I shall not attempt to reproduce here, change is impossible and thus time itself is illusory. A related peculiarity was Melissus' argument that space (or "void") cannot exist because, by definition, nothingness does

[100] The earlier philosopher, Thales of Miletus, regarded by Aristotle as the very first Greek philosopher, was a thoroughgoing animist, or "hylozoist"—a term which means he believed that all matter, animate and inanimate, is alive and willful, or even possibly conscious.

not exist. Despite these strange defects that contradict everyday experience, the writings of the Eleatic tradition are the earliest written record we have of a truly disenchanted materialism.

A generation or so later, a philosopher by the name of Anaxagoras published a work expounding among other things a materialist philosophy intended to revise Parmenides' peculiar view on change. Anaxagoras claimed that elemental materials come in many kinds that are infinitely divisible and, by being mixed in various proportions, these elemental materials compose the many tangible materials we know—the perceptible nature of a thing being established by whatever elemental material that thing contains in the highest proportion. For Anaxagoras change consists in the various mechanical interactions that adjust those proportions (Barnes 1982).

The early atomists Leucippus[101] and Democritus added to Anaxagoras' elemental picture, firstly, that substances are composed of basic, indivisible and immutable objects—*atoms*—of which there are an infinite number, sortable into many, many types, and, secondly (contra Melissus) that there is an enormous space—*the void*—in which these many atoms move about and interact with one another. On Democritus' account, the atoms were outfitted with various ways of interacting; some had hooks and eyelets, or balls and sockets, in order to connect to one another as solids, others were smooth and slippery, accounting for the fluidity of liquids. Everything that occurs, on the atomist account, can be attributed to the mechanical interactions of various kinds of atoms (Barnes 1982).

These ancient materialist views were important to teleological thinking not just for what they said about the material constitution of the world, but for what they didn't say about the causal organization of the world: Since the cause of all the substances and objects and their arrangements in the world can purportedly be accounted for fully by the mechanical natures and motions of atoms within the void, atomism avoids making animistic and theological claims about the behavior and

[101] It is apparently unclear whether Leucippus was a real person, but of interest here is only the work attributed to him.

organization of things. In ancient atomism, matter is not alive and nothing is to be accounted for in terms of reasons or ends or goal-directedness. Nothing is *for* anything.

Platonic Teleology

The philosophers of the Athenian school accepted the Pre-Socratic materialism in part, but found that there was much more to be explained in our world than merely the physical behavior of materials.

In Plato's dialogue the *Phaedo,* Socrates describes his disappointment upon reading Anaxagoras' book. He had hoped to find an explanation of the way the world is arranged in terms of what was "best" and he said that these are answers he would have paid dearly for, and yet he found nothing of the sort. The notion of "best", here, implies that Socrates wanted the arrangements of the world to be explained as not just behaving the way they do in terms of material causes, but doing so for some *reason* that someone thinks is good—that serves some normative purpose. He said, if given in terms of what was best, "I should be satisfied with the explanation given, and not want any other sort of cause," while a description of the world in terms only of materialist mechanisms leaves some things unexplained.

Socrates explains the difference between material mechanism and his notion of "the best" with an analogy, describing the reasons he sits in a jail cell rather than "playing truant" and trying to escape. First, he gives what he thinks Anaxagoras and the atomists would describe as the cause of his sitting there. They would say, he suggests:

> I sit here because my body is made up of bones and muscles; and the bones, as he would say, are hard and have joints which divide them, and the muscles are elastic,

> and they cover the bones, which have also a covering or environment of flesh and skin which contains them; and as the bones are lifted at their joints by the contraction or relaxation of the muscles, I am able to bend my limbs, and this is why I am sitting here in a curved posture. (*Phaedo* 98)

He contrasts this material cause with what he claims is the true cause of his sitting there.

> The Athenians have thought fit to condemn me, and accordingly I have thought it better and more right to remain here and undergo my sentence . . . It may be said, indeed, that without bones and muscles and the other parts of the body I cannot execute my purposes. But to say that I do as I do because of them, and that this is the way in which mind acts, and not from the choice of the best, is a very careless and idle mode of speaking (*Phaedo* 99).

The true cause of his behavior, Socrates believes, is an end, a purpose . . . a conscious choice of what is best which directs his behavior. He also tells us that the orderly arrangement of the cosmos is, somehow, what is best, and that it is similarly directed towards some ends, though he admits he does not have an explanation of quite what ends or whose those might be, nor how they might effect that arrangement.

In the *Timaeus*, however, Plato's eponymous character placates Socrates with just such an explanation. While Timaeus tells us that the structure of the universe follows orderly, mechanistic rules, nonetheless this is because a *creator* intended it . . . for the best.

> Wherefore also finding the whole visible sphere not at rest, but moving in an

> irregular and disorderly fashion, out of disorder [the creator] brought order, considering that this was in every way better than the other (*Timaeus* 30).

With this explanation, Timaeus and Socrates accept both material causes and ends as influencing the way the world is—they accept both "how come" and "what for" reasons—though, as we soon find out, one is more influential than the other.

> The creation is mixed, being made up of necessity and mind. Mind, the ruling power, persuaded necessity to bring the greater part of created things to perfection, and thus and after this manner in the beginning, when the influence of reason got the better of necessity, the universe was created. But if a person will truly tell of the way in which the work was accomplished, he must include the other influence of the variable cause as well. (*Timaeus* 48)

According to Plato's writings, then, the reasons why the world is the way it is fall into two broad categories that work together. "Necessity" is the mechanical, physical description of *how* something comes to be—the material cause. "Mind", which he finds to be more dominant (*i.e.* "reason got the better of necessity"), is the description of *why* something comes to be—the final cause, as it has come to be known[102]. This kind of reason has its source in the mind of a creator and it is that creator's intentions and aesthetic principles that explain why the world is the way it is. And so Plato rejects the purity of mechanism of the atomists and, instead, reintroduces ends into our worldview. They are, however, no longer the ends of animistic spirits but instead those of a god[103].

[102] The term "final" here does not mean "last", in the sense for example that there is some order or priority to the causes; it refers instead to the fact that this kind of cause (of a phenomenon) is an *end* or a result.
[103] The term he uses is "demiurge", which literally means "craftsman".

Aristotelian Teleology

Aristotle inherited from Plato[104] the idea that the physical explanation of many things leaves us wanting for a further, finalistic explanation of some sort. Much as the Platonic account is framed in opposition to Anaxagoras' materialism, so Aristotle framed his own in opposition to Democritus' atomism.

> Democritus, however, neglecting the final cause, reduces to necessity all the operations of nature. Now they are necessary, it is true, but yet they are for a final cause and for the sake of what is best in each case." (*Generation of Animals* V.8)

While both Plato and Aristotle found material explanations to overlook final causes, that is really where the similarities between their teleological accounts end.

In both his *Physics* (II.3) and his *Metaphysics* (V.2), Aristotle distinguishes a total of four types of causes—or categories of reasons—that can be used to describe why the world is the way it is. Three of these, the *material*, *formal*, and *efficient* causes are all further subdivisions of Plato's "necessity" and, taken together, they align more or less with the modern physical account of causation, as described in the previous chapter. Aristotle's *final* cause aligns with Plato's in describing the role of ends in satisfying our explanatory curiosity but, while Plato attributed finality to "mind", for Aristotle final causes are the products of nature alone; the representation that occurs in a mind is inessential.

[104] It is conventional to refer to the views given in Plato's writings as being his, even though Plato, by writing in the form of dialogues, attributes the ideas to his teacher, Socrates, and other contemporaneous thinkers and students of Socrates (such as Timaeus). I'll stick with the convention.

> This is most obvious in the animals other than man: they make things neither by art nor after inquiry or deliberation. Wherefore people discuss whether it is by intelligence or by some other faculty that these creatures work, spiders, ants, and the like [. . .] It is absurd to suppose that purpose is not present because we do not observe the agent deliberating. (*Physics* II.8)

Aristotle sees final causes as being rooted not in the mind of a designer or any other agent but instead in "natures"—reasons that are embodied in the structures of items and organisms themselves. With this idea, he introduces the notion of *function* to teleology—each part of a body, he says, is made, and thus has it in its nature, to serve some partial end (*On the Parts of Animals* I.5) and he gives descriptions, throughout some of his works, of the various parts of anatomy and the ends that they serve. The phrase that best captures this aspect of Aristotelian finality, and which he frequently repeats is "that for the sake of which"—the end which justifies the existence of an item or its behavior. So, health is that for the sake of which there is walking and the windpipe is that for the sake of which there exists the neck ("For [the neck] acts as a defense to [the windpipe] and to the esophagus, encircling them and keeping them from injury", *On The Parts of Animals* IV.10). Translating these ideas into today's vernacular, we might say that a function of walking is maintaining or improving health and a function of the neck is defending the windpipe and the esophagus.

Ultimately, the kinds of items that Aristotle attributes a "that for the sake of which" type of finality to can be categorized into four main groups: Actions performed for the sake of something, objects which exist for the sake of something, processes in organisms which occur for the sake of the organism; and parts of organisms which exist for the sake of the organism (Charles 1995; see also Ariew 2002). The idea that these categories of items have functions has been popular again in

recent decades, though writers in the modern tradition hesitate to refer to such things as being end-directed.

A Recap of Ancient Teleology

While "the *prima facie* teleological characteristics"[105] of the world inspire animistic beliefs and teleological folk-theories in all of us, it was Democritus, Plato, and Aristotle who first set the agenda for Western teleological thinking by distilling into writing the three incommensurable, central hypotheses about purposiveness that would go on to be debated for millennia. Each is superficially plausible but the debate has raged endlessly primarily because each is also, in one way or another, unable to fully satisfy our curiosity. Let's briefly review all three, before moving on to see how that debate has since unfolded.

Democritus gives us the atomistic viewpoint that there are no ends in our world. On this account, teleology is a gratuitous impression irrelevant to explanations of what is, ultimately, a material world. While atomism has been updated heavily in the intervening millennia, modern materialism is in broad detail similar to the Democritean and Anaxagoran versions, and so something like the teleology-free account of the ancient atomists also holds sway with a majority of today's scientists.

Aristotle suggests that there is a kind of immanent or intrinsic end-directedness in both organisms and artifacts that underpins the arrangements of both the human and the biological worlds, accounting for the way items and behaviors can be described in terms of what they are for. This view has gone in and out of vogue over the centuries but, for lack of a naturalistic explanation

[105] This phrase comes from Bedau and Cleland (2010) who intend to emphasize the fact that it isn't, or shouldn't be, controversial that we *observe* goal-directedness in the world; the controversy should only surround our ideas of what accounts for those observations.

of Aristotle's natures, it is largely out of favor today, except in terms of modern function theory, which can be thought of as Aristotelian, though it typically eschews final causation and end-directedness.

Plato offers an account of intentional teleology with both human and deistic expressions that has sat well with creationists and theologians throughout history, and even today. From a certain angle, this account seems to be simpler than Aristotle's—if organisms are seen as a god's artifacts, then all the things we observe that appear to be arranged for some end stem simply from their having been created by some intentional agent. Ends are simply intentions or desires. Since, of the three ancient accounts, Plato's had the most influence over the next two thousand years, let's look at those developments now.

C. The Teleological Argument

For it is a Sign a Man is a wilful, perverse Atheist, that will impute so glorious a Work, as the Creation is, to any Thing, yea, a mere Nothing (as Chance is) rather than to God.

—William Derham (1713:328)

The general account voiced by Socrates and Timaeus remained a common argument for the existence of a creator throughout later antiquity and the middle ages, being repeated, for example, by St. Augustine (in his *City of God*), St. Thomas Aquinas (in his *Summa Theologiae*), and John Ray (1691) of the Christian faith, and by the philosophers Averroes (in his *Tahafut Al-Tahafut)* and Al-Ghazali (in his *al-Hikmah fi makhluqat Allah*) of the Islamic faith, as well as by Maimonides of the Jewish faith (in his *Moreh Nevukhim*). It eventually became known as the "teleological argument" or, more commonly, the "argument from design"[106] since it starts with the observation that some things in the natural world appear to be designed, and then reasons by analogy to the existence of a designer—typically some god or other. On this view, the purposes we find in the natural world all derive somehow from a creator's intentions the same way that the purposes we find in human artifacts derive from our own intentions.

The now-classic form of the argument is what's known as the watchmaker analogy, which shows up first in Cicero's writing, dating from 45 BCE.

> When we see something moved by machinery, like an orrery or clock or many other such things, we do not doubt that these contrivances are the work of reason; when

[106] Paley (1802) appears to be the first to use the name "argument from design".

> therefore we behold the whole compass of the heaven moving with revolutions of marvellous velocity and executing with perfect regularity the annual changes of the seasons with absolute safety and security for all things, how can we doubt that all this is effected not merely by reason, but by a reason that is transcendent and divine. (*De Natura Deorum*)[107]

The analogy has been transmitted mostly unchanged over the centuries and today it is still a mainstay for proponents of intelligent design, but it was William Paley who, in the early nineteenth century, provided what is now the most-cited version of the argument. Paley's argument begins by questioning what we might think if we discovered a pocket watch in a heathland. The assumption he foists upon us (not unreasonably) is that the watch had not simply come to be there by geological forces as a rock might have, but had instead been created by a designer and somehow left in the heath.

> This mechanism being observed . . . the inference, we think, is inevitable, that the watch must have had a maker: that there must have existed, at some time, and at some place or other, an artificer or artificers who formed it for the purpose which we find it actually to answer; who comprehended its construction, and designed its use. (Paley 1802)

[107] Another more oft-quoted passage from the same work is the following: "When you see a statue or a painting, you recognize the exercise of art; when you observe from a distance the course of a ship, you do not hesitate to assume that its motion is guided by reason and by art; when you look at a sun-dial or a water-clock, you infer that it tells the time by art and not by chance; how then can it be consistent to suppose that the world, which includes both the works of art in question, the craftsmen who made them, and everything else besides, can be devoid of purpose and of reason?" (*De Natura Deorum*).

Paley then suggests that if we were to discover not a watch but a *self-replicating* watch, such a discovery would not change our convictions but only "increase, beyond measure, our admiration of the skill, which had been employed in the formation of the machine" (ibid). We would be convinced beyond a doubt that such a complex machine had to have been the creation of an artificer. And he concludes by suggesting it is absurd to imagine that a self-replicating machine could have come to exist by any method other than art and skill, stating at the last, "Yet this is atheism" (ibid). We are meant to accept that biological organisms are equivalent to such self-replicating watches and so we must recognize that their forms could only have sprung from the mind of a magnificent designer.

Modern proponents of evolutionary theory have often argued against Paley on the basic premise that, while the watchmaker analogy makes it clear that design in biology needs to be explained, it doesn't justify the conclusion that only a divine designer could explain it. Not only would the divine designer itself be a magnificent machine still requiring explanation (making Paley's argument regressive) but we also have a very elegant (and non-regressive) alternative explanation for the existence of organisms: natural selection. Descent with modification is a natural form of design that requires no foresight or intention in order to create complex biological mechanisms (Darwin 1959; Dawkins 1986, 2006; Dennett 1995)[108].

I am of course loath to add apparent fuel to the creationist's furnace. It is, however, worth examining a hole left behind by the considerations just mentioned: Ruling out divine *design*, does not necessarily rule out all divine *teleology*. Another teleological argument—call it the "argument from vitality"—can at least be imagined. That is to say, the creationist may happily accept evolution by

[108] Although I find the points just mentioned to be the most convincing arguments against the watchmaker analogy, others have also been given (see *e.g.* Hume 1779; Mill 1874; Salmon 1978; Richerson and Boyd 1995; Voltaire 1734).

natural selection[109] as a replacement for divine design in creating the complexity of biological mechanisms and yet they may still wonder why it is that the items created by that process are not only *intricate* (as design might create), but also *purposeful* and *vitalistic*. Could not a god have set into motion the process of evolution and yet also seeded the first organisms with a special kind of purposiveness or a spark of life that is then passed down the reproductive and evolutionary line?

Gánti (2003) argues, as will I, that while Darwinian evolution explains the history and diversity of life, it indeed leaves its vitality and purposiveness entirely unexplained (see also Mayr 1997; Oparin 1964; and Schrödinger 1967). And in fact, near the end of Charles Darwin's life, in a letter to his friend T.H. Farrer, Darwin wrote "[I]f we consider the whole universe, the mind refuses to look at it as the outcome of chance—that is, without design or purpose. The whole question seems to me insoluble" (Darwin and Seward 1903). The very man who eradicated intentional design from our understanding of the evolution and adaptation of organisms was still mystified by the purposiveness found in the world. But is the argument from vitality—reasoning to the existence of a divine life-giver—the answer to these mysteries? I don't think it is.

As we saw, there were two parts to the modern dismissal of the argument from design. The first—that it is regressive—shows that nothing is answered by the divine hypothesis. The second—that there is a plausible and non-regressive alternative—shows that we are not stuck in a position where we need a divine hypothesis. We have a better answer. The same two types of contention can also dismiss the argument from vitality. First, the notion that a (living, purposive) god makes biological items alive and purposive is just as regressive as is the argument from design. It invites the question: what makes the god itself alive and purposive in the first place? As with the argument from design, the divine answer is simply not a good answer because, while it tells a story, it entirely

[109] Indeed, it has been suggested that had Paley lived long enough to learn of Darwin's breakthrough, that he would have embraced evolution by natural selection and found it compatible with his theistic beliefs, much as Darwin himself and a number of his contemporaries did (Shapiro 2009).

fails in terms of explaining the phenomenon for which we are seeking an explanation. Second, a god is only one hypothesis of possibly many. And, while it is true that we don't yet have a strong alternative explanation to account for vitality and purposiveness, that doesn't mean one can't be found. As we'll see later, theorists have been attempting to find just such an alternative.

D. The Scientific Revolution

One of [Darwin's] greatest accomplishments was to bring the teleological aspects of nature into the realm of science. He substituted a scientific teleology for a theological one. The teleology of nature could now be explained, at least in principle, as the result of natural laws manifested in natural processes, without recourse to an external creator or to spiritual or nonmaterial forces. At that point biology came into maturity as a science.

—Francisco Ayala (1968)

In the seventeenth century, for the first time since the period of the ancient atomists, a new philosophical rejection of teleology began to gain traction. As physics developed, it became clearer and clearer that there was no place for ends in the mechanics of physical nature; and, on the assumption that biology and human affairs were all parts of the physical world, philosophers began to push the notion that, even in these realms, ends may be superfluous.

The start of this rejection came when Francis Bacon published his *Novum Organum* in 1620, therein denouncing final causes in favor of material, formal and efficient causes, and going so far as to say "Of these [Aristotle's four causes], however, [the final cause] is so far from being beneficial, that it even corrupts the sciences, except in the intercourse of man with man".

A similar sentiment is implied in Galileo's work from around the same time. His 1616 *Discourse on the Tides* and his posthumous *De Motu* ("On Motion") both speak of "fundamental causes", "true causes" and "primary causes", every mention of which is based in physical concepts—material, formal, and efficient causes—rather than reasons or ends. In *Il Saggiatore* ("The Assayer", 1623) Galileo illustrated his vision for natural philosophy (which at the time meant

physics) to be performed purely in the language of mathematics, notably excluding any mention of ends or reasons. Galileo's physics doesn't ask what the behaviors of objects are *for*.

Eventually, as the scientific revolution got into full swing in the middle of the century, philosophers, too, took up the fight. René Descartes was convinced that biological systems are just complex fluid mechanical (physical) machines and that mechanistic, physical explanations can account for their behavior. With a tone similar to Bacon's, Descartes explicitly and repeatedly renounced final causes (*Meditationes de Prima Philosophia*, 1641; *Principia Philosophiae,* 1644; see also, de la Mettrie 1748)[110].

Benedict de Spinoza made a still stronger statement, in the appendix to book I of *The Ethics*, where he argued that "nature has no particular goal in view, and that final causes are mere human figments". Spinoza meant to account for the illusion of teleology and so blamed the appearance of purposes in nature on our tendency to (falsely) believe that if we did not create things for our own purposes, the gods would have done so for us. He cemented his position with what is possibly the first argument against finality in terms of backwards causation. He says "That which is really a cause it considers as an effect, and vice versâ: it makes that which is by nature first to be last." (1677).

Newton and Laplace

It is said that humanity made good in full on the scientific revolution in 1687, when Isaac Newton published his *Philosophiae Naturalis Principia Mathematica*, but it was here that the scientific rejection of teleology actually saw some resistance. Although Newton's publication may have laid the groundwork for the demise of teleology, he himself was unable to accept that there were no ends in physics. It wasn't until the turn of the nineteenth century that the work of Pierre-Simon Laplace

[110] According to Simmons' (2001) interpretation, despite Descartes' arguing against finality, he may have held some "latent teleology" in the ways he sometimes discussed the body and its role in survival.

appeared to tie up the loose ends that Newton noticed, at last convincing many that there really was no place for final causation in physics.

Like Galileo, Newton argued for a practical, natural style of reasoning in science, based largely in mathematics. At the start of Book Three of the *Principia*, in a section entitled "Hypotheses" (in later editions, "Rules for Reasoning in Natural Philosophy"), he stated the now famous Hypothesis I.

> Hypothesis I: Causas rerum naturalium non plures admitti debere, qu`am quæ & vera sint & earum Phænomenis explicandis sufficiunt. Natura enim simplex est & rerum causis superfluis non luxuriat.
>
> English Translation: We ought admit no more causes of natural things than those that are both true and sufficient to explain their phenomena. Nature is in fact simple and doesn't luxuriate in unnecessary causes.

In other words, Hypothesis I asks us to take an explanation as complete if it is fully predictive of the phenomenon it is meant to explain. It is through this dictum, still widely regarded and followed by scientists today, that teleology might have been properly banished from physical (though not biological) science[111]. That is, if, for instance, the material cause of gravitation can explain the motions of bodies, then we should not look further for a final cause in the motion of bodies, since *there is nothing left to be explained.*

[111] In his second edition, published in 1713, Newton added his famous dictum, "hypotheses non fingo" or, in English "I do not make (or contrive) hypotheses". Though often quoted with this brevity, the claim is difficult to comprehend as such, considering that Newton regularly worked from and sometimes published hypotheses. In its context, though, we find that what Newton meant by the dictum is that explanations of physical phenomena that are based in observations do not require us to hypothesize further "deeper reasons" for their occurrence. In short, he was arguing against animistic, occult, vitalistic, and teleological explanations for physical phenomena.

But Newton himself was unconvinced that his mechanistic physics was complete. He felt that the motions of bodies were not fully explained and he was puzzled still as to the nature of gravitation. In a letter to Richard Bentley, he confessed to believing that *agency* guides gravitation.

> That gravity should be innate inherent & essential to matter so that one body may act upon another at a distance through a vacuum without the mediation of any thing else by & through which their action or force may be conveyed from one to another is to me so great an absurdity that I believe no man who has in philosophical matters any competent faculty of thinking can ever fall into it. Gravity must be caused by an agent acting constantly according to certain laws, but whether this agent be material or immaterial is a question I have left to the consideration of my readers (1692b).

Combined with both his theological beliefs and the regularity of the solar system, this convinced him of something much like the watchmaker analogy. In another letter to Bentley, he suggested that there is an intelligent designer who set up the world just so, and who occasionally intervenes in order to keep things running.

> To make this systeme therefore with all its motions, required a Cause which understood & compared together the quantities of matter in the several bodies of the Sun & Planets & the gravitating powers resulting from thence, the several distances of the primary Planets from the Sun & secondary ones from Saturn Iupiter & the earth, & the velocities with which these Planets could revolve at those distances about those quantities of matter in the central bodies. And to compare & adjust all

> these things together in so great a variety of bodies argues that cause to be not blind & fortuitous, but very well skilled in Mechanicks & Geometry. (1692a)

The argument is that the precision required to arrange the orbits of the planets such that gravity would then keep them orbiting is an astonishing feat requiring a careful architect.

In addition, Newton noticed that the orbital speeds of Jupiter and Saturn were such that Jupiter's orbit would shrink and Saturn's expand, causing the solar system to eventually fall apart. He assumed that this could not happen, however, and so conjectured (1704) that the creator, so "skilled in Mechanicks and Geometry", had arranged the eccentric orbits of the comets such that they would pass by the planets occasionally, thus restabilizing their orbits (see also Amundson 1996; Ariew 2002). We now know that the comets play no such role and, while there is some stability in the planetary orbits, the solar system is in fact chaotic and those orbits will, over longer time spans, vary widely and eventually decay catastrophically[112]. The extreme regularity that Newton attributed to the design or intervention of a god simply doesn't exist (Walsh 2009; Walsh *et al.* 2011).

Ultimately, despite his commitment in the *Principia* to a mathematical, material physics without unnecessary causes, Newton was unable to believe that the behaviors of celestial bodies could be explained without agentive intervention. He still believed in a kind of Platonic creator whose purposive hand was behind it all. According to Ariew (2002, 2007; see also Amundson 1996), it was Laplace, who ultimately closed the book on teleology in physics. Laplace ushered in this conclusion, in the first place, with his early explanations of the formation of the solar system (1796) which precluded an architect, and his (not quite right) discovery that the orbits of Jupiter and Saturn would self-correct, which allayed Newtonian concerns about the necessity of a designer and maintainer for continued stability of the solar system. Laplace's divinity-free version of Newtonian

[112] As it turns out, modern models indicate that the interactions between the planets are unlikely to catastrophically disturb one another's orbits for at least many millions, if not billions, of years (Hayes 2007).

physics was further fortified as he developed his (1814) philosophy of determinism, which ruled out any design or intervention in the motions of matter. It is at this point that western culture finally arrived at the deterministic materialism that underlies the modern scientific worldview, making space for the development of atomism and 20th-century particle physics, but also laying the foundations for the largely anti-teleological bias that persists today.

Kant's Biological Teleology

Much of Kant's (1790) *Critique of the Power of Judgment* is an attempt to account for the goal-directedness found in biology. I must admit, I find it difficult to confidently interpret Kant's teleological work as a whole, partly because of his style of writing, but also because he offers both Platonic and Aristotelian aspects, which seem to me in conflict with one another. However, the Aristotelian branch of his work has notable innovations as compared to any writer before him, including his treatment of artifacts in terms of their service and his treatment of organisms in terms of their self-directed causality. Within these two ideas lies an important germ of the account I'll describe later.

The more Aristotelian aspect of Kant's account shows up in his discussion of "the special character of things as natural ends" (§64) wherein, like Aristotle, he discusses an innate or immanent type of purposiveness that exists in organisms. Kant says "provisionally . . . a thing exists as a natural end if it is cause and effect of itself" (1790: 243) and he enumerates three ways in which a thing might be both cause and effect of itself. First, something might be a replicator—a tree, he says, generates another tree and so is cause and effect of itself *as a species*. Second, the tree also generates itself through *growth and development*. And, third, the tree's various parts have a reciprocal dependency upon one another such that each contributes to the preservation of the others and, all

together, the tree's parts maintain preservation of itself—a process he calls "*self-help*" (ibid: 244) and which essentially prefigures Maturana and Varela's (1973) theory of autopoiesis by nearly two centuries (see also Cuvier 1798). Along with the notion of "cause and effect of itself", two of these three examples—replication and autopoiesis—will be central to the account of teleology I will describe in part II.

For Kant, the purposiveness of artifacts is not internal, the way it is for organisms; it is, he says, "an entirely different concept". Principally he speaks of the way organisms use one another to serve their own needs (for instance the animal kingdom gets nourishment by feeding on the vegetable kingdom) but the same reasoning applies to artifacts as well which get their purposes from their organismic users. He says, "By external [or relative] purposiveness I mean that in which one thing in nature serves another as the means to an end" (ibid: §82, p. 293). This, too, is important to the account I will describe in part II.

Darwin, on Teleology

Darwin's (1859) work on evolution is rivaled in its contribution to biology only perhaps by the discovery of the structure and nature of DNA (Watson and Crick 1953). As an explanation of both the diversity of life and of its astonishingly designed character, the theory of natural selection was a monumental achievement for our understanding of the phenomena of life. And yet it has been challenging for thinkers over the past century and a half to figure out just what teleological conclusions we can draw from Darwin's theory. Ernst Mayr cites conflicting interpretations:

> David Hull (1973) has recently stated that "evolutionary theory did away with teleology, and that is that," yet, a few years earlier MacLeod (1957) had pronounced

> "what is most challenging about Darwin, is his reintroduction of purpose into the natural world." Obviously the two authors must mean very different things. (Mayr 1974)

In the same vein, Dennett (2014) points out that Karl Marx seems to agree with both Hull and MacLeod. Here is Marx, writing just two years after the publication of *On the Origin of Species*:

> It is here that, for the first time, "teleology" in natural sciences is not only dealt a death blow but its rational meaning is empirically explained. (Marx 1861)

Dennett suggests that Marx is equivocating; his "death blow" appears to banish teleology from the natural sciences while at the same time his empirical explanation of its "rational meaning" retains teleology by replacing various cosmic, animist, vitalist, and theological views with a natural Darwinian explanation of some sort (Dennett 2014).

Darwin himself did not know quite what to make of teleology either. He'd read Paley's work during his time studying at Christ's College in Cambridge and, at the time, apparently was convinced by the argument from design (Darwin 1860:258). Over the decades to come, however, he came to realize that his own theory of evolution by natural selection rendered a designer in biology unnecessarily redundant.

Yet, abandoning the argument from design did not give Darwin anti-teleological leanings. As James Lennox writes, Darwin wrote openly of final causes throughout his *Species Notebooks* (Barrett *et al.* 1987), using the term synonymously with claims about what a trait is *for*, and he continued this practice in *On the Origin of Species* (Darwin 1964: 216, 435, 448, as cited in Lennox 1993).

> More importantly than his bare use of the term, however, are his consistent arguments that natural selection acts for the good of each being, and that its products are present for various functions, purposes and ends (Darwin 1964, 149, 152, 224, 237, 451). As John Beatty (1990, 127) and Ernst Mayr (1988, 241) point out, that was all Albert von Kölliker, a contemporary of Darwin's and a critic of teleology, needed to be assured that Darwin was a teleologist. (Lennox 1993)

We also know, from a series of exchanges between Darwin and the Harvard University botanist, Asa Gray, that even after the publication of his theory, Darwin appeared entirely positive about teleology while largely[113] negative about creation-style designedness. In a review of Darwinian thinking in *Nature*, Gray wrote:

> Apropos to these papers, which furnish excellent illustrations of it, let us recognise Darwin's great service to Natural Science in bringing back to it Teleology: so that, instead of Morphology *versus* Teleology, we shall have Morphology wedded to Teleology. (Gray 1874: 81)

Darwin agreed entirely, writing in a letter to Gray, "What you say about Teleology pleases me especially, and I do not think any one else has ever noticed the point" (Darwin 1959: 367).

[113] There are moments where Darwin shows some conflict about design. In particular, he writes to Gray, "On the other hand I cannot anyhow be contented to view this wonderful universe & especially the nature of man, & to conclude that everything is the result of brute force. I am inclined to look at everything as resulting from designed laws, with the details, whether good or bad, left to the working out of what we may call chance. Not that this notion *at all* satisfies me . . . " (Darwin 1860:224) and, later, "[i]f anything is designed, certainly Man must be; yet I cannot admit that man's rudimentary mammae . . . & pug-nose were designed . . .. I am in thick mud;--the orthodox would say in fetid abominable mud" (Darwin 1861:369).

Perhaps the interpretation that best makes sense of both the quote from Marx and Darwin's own ostensible equivocation is that Platonic, theological, teleology—Paley's argument from design—is dealt a deathblow while Darwinian processes can account for—or at least be consistent with—the reasons biological items seem to be functional or serve ends, in the immanent, Aristotelian, sense.

I think this is the case. However, there are at least two manners in which the post-Darwinian thinker might conceive of an intrinsic, Aristotelian end-directedness. One way is for natural selection to have furnished, and to be the full explanation of, those ends. Dennett puts it this way: "[Darwinian] processes brought purposes and reasons into existence the same way they brought color vision (and hence colors) into existence: gradually." (Dennett 2014: 49). This is also the view that Marx and Darwin appear to have accepted and that numerous other authors in recent decades have espoused (see Ariew 2002; Lennox 1993; and the various contributors to the selected effects theory in Chapter V). It is intuitive but I think it is ultimately an incomplete account.

The other way to conceive of a post-Darwinian, Aristotelian end is to see items as having some kind of *structural* rather than *historical* nature that accounts for their end-directedness. This is more closely aligned both with the vitalist position (discussion of which we'll take up shortly) and the cybernetic theory (see above) and with the causal role and replication disposition theories that we'll look at in Chapter V.

Darwin's theory nullifies the argument from design, but it doesn't eliminate teleology (see also Butler 1879, 1880). In fact, I'll try to show later that Darwinian natural selection *assumes* teleology rather than accounting for it. As we see from Darwin's writing and from the writings of legions of evolutionary biologists since, discussion of the evolved traits of organisms is best couched in terms of the teleological question of what the trait is for, what purpose it serves, or what function it has (see Allen and Bekoff 1995a; Beckner 1969). But this indicates a logic reversed from that of

natural selection creating purposes: If a trait or new mutation turns out to be good for something, if it has or serves a purpose in the current environment, then it becomes a candidate for selection to thereby *preserve* its already teleological nature.

Evolution as Teleological

In the years following Darwin, it had sometimes been proposed that the process of evolution itself is, or is subject to, some kind of teleological or directive force by which the various biological items produced are progressively more advanced or lie upon an intended path (Berg 1926; Bergson 1911; Campbell 1985; Teilhard 1955; Eimer 1890, 1897; Haacke 1983; Kellogg 1907; Lovejoy 1936; Osborn 1934). Along with some versions of this progressionism[114], often comes the homocentric belief that, if humankind is not the pinnacle of the progression then it is certainly well along a singular directed path toward some ultimate end result that is perhaps similar to us.

Depending on the version of the theory, the teleological force supposedly guiding progressionist evolution is taken either to be divine or somehow internal to organisms or their DNA, and evolution by progressionism has sometimes been taken to be an alternative theory to evolution by natural selection. Darwin's aforementioned colleague, Asa Gray, expressed a creationist version of the belief (1963a). Darwin, however, respectfully disagreed.

> However much we may wish it, we can hardly follow Professor Asa Gray in his belief that "variation has been led along certain beneficial lines," like a stream "along definite and useful lines of irrigation." (Darwin 1896:428)

[114] Also called orthogenesis and, sometimes, autogenesis.

More recently, Nick Lane (2010) published an insightful examination of modern origin-of-life theories and of some of the most powerfully earth-changing biological paradigms to be developed by evolution since. It is a delightful book to read but while Lane does not give an explicitly progressionist thesis, his title, *Life Ascending*, certainly implies one. The progression he charts, from acellular metabolic processes to the development of replicating DNA, photosynthesis, and cell-walls to multicellularity and sexual reproduction and then to motility, vision, warm-bloodedness and finally the capacity for conscious reflection[115], also raises an obvious question: what accounts for the seemingly end-directed nature of evolutionary change?

The answer to that question comes quite simply from the fact that we might tend to ignore the backwards and sideways movements, as well as the fatal missteps, that evolution constantly makes in its diffusive explorations[116]. When natural selection was a nascent theory, and the tree of life was first being sketched, it was easy to imagine that evolution proceeded smoothly from simpler to more sophisticated "higher" organisms. However, we now have heaps upon heaps of genetic and paleontological taxonomic evidence (as well as logical reasoning) that suggest evolution simply explores all directions, with many of its trials being unviable and not getting out of the blocks, and many others moving in directions not considered to be so-called progress. The majority of these trials are largely invisible to us because, due to their unsuccessful nature, they leave far fewer records to be examined[117]. As biologists in the twentieth century came to this understanding, the progressionist viewpoint simply lost its credibility (Mayr 1992; O'Grady 1984; Simpson 1949, 1953, 1964; Weismann 1909). While goal-directedness and function in organisms and artifacts are patterns

[115] Lane's last chapter is on the topic of death and, as he tells it, the way we humans experience (non-traumatic, senescent) death is as much a result of evolution as are the rest of the topics he covers (see also Medawar 1952; Williams 1957). In terms of Lane's progression, however, it is something of an outlier that happens to be last in line simply because it couldn't have been first.

[116] These evolutionary changes are not *truly* backwards or sideways, since there is no favored direction, just as the progressive-seeming moves are not forward. Evolution simply explores any and all directions, through random mutation in genetic space.

[117] This notion is now called "survivorship bias", following Abraham Wald's exposure of an error in reasoning during analysis of damage patterns on surviving warplanes (Wald 1943; see also Taleb 2001, 2007).

worth investigation, there is no good reason to believe that the process of evolution itself is driven by any particular goal.

Final Thoughts on Final Causation

Belief in final causation, in a variety of forms, has come and gone over and over again as science and philosophy have advanced, but it has never disappeared entirely. As we've just seen, even in the mid to late nineteenth century, there were still those, like Darwin, who believed in an immanent, Aristotelian teleology; those, such as Gray and followers of Paley, who believed in a theological, Platonic teleology; those, such as Wilhelm Haacke and Theodor Eimer, who believed in teleologically progressive evolution; and even those, such as Hermann von Helmholtz, Emil du Bois-Reymond, Karl Ludwig, and Ernst Brücke, who took a firmly materialist and anti-teleological, Democritean stance[118]. The tension between, on the one hand, the developing materialist scientific worldview and, on the other hand, the existence of end-directed patterns that call out for teleological answers was never resolved, and still isn't today. We'll see a bit more of this tension when we review the history of vitalism.

[118] Helmholtz, du Bois-Reymond, Ludwig, and Brücke are famous for making a pact, swearing to denounce any non-materialist principles in biology, in spite of the vitalist teachings of their mentor, Johannes Müller. The text of the pact, known as "the Reymond-Brücke oath", is: "No other forces than the common physical-chemical ones are active within the organism. In those cases which cannot at the time be explained by these forces one has either to find the specific way or form of their action by means of the physical mathematical method, or to assume new forces equal in dignity to the physical-chemical forces inherent in matter, reducible to the force of attraction and repulsion." (trans. Bernfeld 1949:171).

E. Vitalism

Every living thing therefore, whether animal or vegetable, owes its vitality to the heat contained within it. From this it must be inferred that this element of heat possesses in itself a vital force that pervades the whole world.

—Cicero (*De Natura Duorem: II.ix.24*)

Only an insufficient acquaintance with the forces of inorganic nature can account for the frequent denial of the existence of a special force in organic beings, and for the ascription to inorganic forces of modes of action which are opposed to their nature and which contradict their laws. . . . In living bodies there is added yet a fourth cause which dominates the force of cohesion and combines the elements in new forms so that they gain new qualities—forms and qualities which do not appear except in the organism.

—Justus von Liebig (1844, as cited in Driesch 1914)

The living and the non-living are, in general, clearly distinguishable, but precisely wherein lies the difference is not easily stated. To declare that a living being has a soul, by virtue of which it is alive, does not advance our knowledge very far. It merely restates the original problem and says tautologically that whatever lives has a principle of life.

—Lester King (1964)

The Aspect of Vitality

A living organism has a strange thermodynamic destiny different from other very similar bundles of matter. One way to put this into stark relief is to think about what happens when we die. Think of a human who has just undergone ventricular fibrillation leading to cardiac arrest. Very little, at first, has changed about the structure of the matter that we think of as that person—its heart has simply stopped beating properly. For a short window of time, it may even be possible to restart the heart and keep the individual alive, but once that opportunity has passed, the thermodynamic future of the bundle of matter that once was that person is very different than had it not experienced the ventricular fibrillation. The tendency toward thermodynamic equilibrium will drain the organized structure of its free energy, and the process of ratcheted braising will dismantle it much more quickly now than before[119]. What is it that makes the same lump of matter at one moment alive, and at the next moment dead? What strange vital force operates in the one and not the other?

For a long time, theorists indeed believed there to be a vital force that inhabited living organisms and that accounted for the difference between the animate and the inanimate parts of our world. This *vitalism* differs from some similar doctrines, including what I have been calling *animism*—the belief that willful spirits inhabit many or most items in our world—and Thales' *hylozoism*—the belief that all matter is alive. Vitalism posits an immaterial substance, force, or energy that animates just the living, and that is meant to account for many of the astonishing behaviors of biology, including development, growth, and reproduction, and more generally the aspect of vitality found throughout the biological world.

[119] Jöns Jacob Berzelius made more or less this same observation in his (1827) *Textbook of Chemistry,* although he didn't yet have the language of thermodynamics to use at the time. See also the below section on Stahl's *ens activum* (King 1964).

There were two great waves of vitalism. Following Hans Driesch (1914), we can call the members of the first wave, comprised of ancient philosophers and physicians, the *early vitalists*; members of the second wave—mainly biologists from the seventeenth to nineteenth century, including Driesch himself—can be called the *later vitalists*.

The early vitalists' concern was primarily to account for the aspect of vitality—for the animacy of biological activity, and for such facts as that of human and animal respiration and heat production. Because these thinkers were broadly interested in the existence and the agentive behavior of living things, on the surface their vitalism may seem somewhat more teleological. The question that the early vitalists meant to answer was some version of, "What accounts for the state of being alive?" and their answer was something along the lines of a soul[120].

The later scientific vitalists were trying to make sense of more specific biological phenomena such as embryological development, growth, self-maintenance, healing and regeneration, reproduction, so-called spontaneous generation[121], animal movement, and heredity. A key part of the mystery in all these phenomena was that that they were for the most part informational and organizational processes. Organisms build themselves, rebuild themselves, and eventually reproduce themselves, passing their own structure down to their descendants, thereby creating non-decreasing (and often increasing) amounts of organization in the world and contradicting the thermodynamic imperative of thermal equilibrium. It didn't make sense to the vitalists that such organizational[122] processes as they observed—such obviously *directed* processes—could occur spontaneously, since such extreme organizing tendency rarely appears in non-biological quarters of the world. There had to be something guiding these processes. Absent the observations and experiments that made possible our modern understanding of cells, cellular processes, and the roles of DNA and other

[120] In fact, if we translate "animacy" from the Latin root "anima" it more or less means *soulfulness*.
[121] The appearance of living organisms, especially molds and bacterial colonies, seemingly out of nowhere (although we now know this appearance to be illusory).
[122] In their time, these scientists wouldn't have said informational.

cellular machinery in protein-synthesis and replication, the only available explanation was that some unseen force guided these processes.

In fact, proposing such a metaphysical force was a laudably rational attempt at explanation. Despite vitalism now looking to us like a magical hypothesis, for its proponents it was rather a way of ruling out magic as an explanation: a vital force of some sort could act as a theoretical promissory note—a thing to be searched for and further understood at a later date—while, in the meantime, one could avoid pretending that the mystery of observed informational, organizational, directed, non-equilibrium processes that characterizes life is illusory. Ernst Mayr recently put it this way:

> It would be ahistorical to ridicule vitalism. When one reads the writings of one of the leading vitalists, like Driesch, one is forced to agree with him that many of the basic problems of biology simply cannot be solved by Cartesian philosophy, in which the organism is considered nothing but a machine (Mayr 2002).

Vitalism has declined in popularity to the point that modern thinkers who seem to even imply such a doctrine are usually written off as dualist cranks. But even today, ridicule of vitalism is a scientific conceit. The aspect of vitality has not been explained. Although the facts that concerned the vitalists—that living processes are informational, organizational, directed, and non-equilibrium—are partly explained by our understanding of the information-bearing properties of DNA, they are also partly still a mystery, and good suggestions to try to account for those facts have only recently begun to appear. We still do not have a general biological principle that can account for the aspect of vitality and the organizational processes found in biological organisms.

Vis Vitalis

Like the animistic, theological, and immanent varieties of teleological thinking that we looked at in the previous section, vitalism too can trace its roots back to the ancient Greek philosophers. Aristotle spoke of a "vital spark", equated more or less with the soul, that was the driver of embryological development in animals (*De Anima; De Generatione Animalium*). Later, the Stoic philosopher Cleanthes modified this idea of a spark, suggesting instead that the sun is the divine giver of all life, and that heat itself is the source of vitality that animates living beings (Cicero, *De Natura Deorum*)[123]. This idea of a life-granting substance, eventually renamed "*vis vitalis"* ("vital force") by Posidonius (b. 135 BCE), remained a central dogma of medical thought for many centuries. And we can trace the idea, from there, as least as far forward as the Persian philosopher, Avicenna (11th century CE), who, following the prolific Roman physician Galen[124], also wrote of an "innate heat" that is produced in the heart, the extinction of which is to be equated with death. Avicenna's text, *The Canon of Medicine*, was reportedly used in teaching at medical schools as late as the seventeenth century[125,126].

[123] One might notice the affinity between Cleanthes' idea and Prigogine's modern thermodynamics of dissipative systems. Under Prigogine's interpretation, it is not necessarily heat but, still, it is ultimately energy that flows originally from the sun that powers the dissipative processes that allow the structures of terrestrial life to form and be maintained despite the second law of thermodynamics.

[124] Galen was a student of both Posidonius' and Aristotle's writings. In his books *On the Natural Faculties*, and *On the Usefulness of the Parts of the Body,* he wrote of an "innate heat" and "vital flame" that operates in living beings.

[125] The 1911 edition of the *Encyclopædia Britannica* reports this fact (as cited in Wikisource, "1911_Encyclopædia_Britannica/Avicenna", n.d.).

[126] Also as mentioned earlier, in the seventeenth century Leibniz came to consider kinetic energy—the transfer of which is heat—as a living motive force, and so he called it *vis viva*, quite similarly to Posidonius' *vis vitalis*.

Ens Activum and Vis Essentialis

As the later vitalists of the seventeenth to nineteenth centuries began working in earnest toward understanding the organizational mysteries of biology (development, growth, reproduction, and so on . . .), an interesting trend developed whereby partial and proximate answers to each biological quandary were discovered, but, rather than fully satisfying our curiosity, these answers instead highlighted deeper organizational questions. As a result, the researchers whose observations had helped them to (partially) solve those mysteries very typically would immediately follow up their answers with deeper vital hypotheses. We can get a coarse sense of what this scientific vitalism looked like by reviewing a pair of early examples of the practice, from the physiologist Georg Ernst Stahl and the embryologist Caspar Friedrich Wolff, who are each representative of the tradition.

Stahl, now often the target of ridicule for both his phlogiston theory and his vitalism, was highly influential for that vitalism during the Age of Enlightenment. Although very much a materialist in his investigations of physiological mechanisms such as circulation, secretion, and excretion, Stahl also became convinced that there was something more that *motivated* the physiological mechanisms of life—a director, behind the scenes, that coordinates and animates the organism. For instance, in noticing as we did above that dead organic bodies will decompose while living organisms do not and, recognizing that there needs to be a reason for this, Stahl proposed, much like the Greek philosophers, that there is an immaterial soul—he called it the *ens activum* or "active being"; the source of vitality—which directs the behaviors of the body and drives the mechanical motions of physiological processes to prevent decomposition (King 1964). When the *ens activum* leaves the body at death, the maintenance machinery of the body is left without an operator, and so decomposition proceeds unchecked.

Wolff, through his microscopic observations of the development of plants and animals, effectively refuted the idea of preformation—the embryological notion that mature organisms simply grew larger from tiny, yet complete, versions of themselves[127]. In place of preformationist thinking he reinstated Aristotle's theory of epigenesis, which is organismic growth and development roughly as we understand it today (Wolff 1759). Although Wolff's detailed observational work made it clear *that* epigenesis occurs, it was unable to explain *how* the process worked or *why* it should be the case. Since Wolff had no way of knowing what organizational structure could possibly control and direct the growth and differentiation of the parts of organisms, he gave in his embryological account an explanation by which seeds and embryos developed into organisms through the action of a metaphysical driving force which he called *vis essentialis*, the "essential force" which acted as an intelligent supervisor coordinating the assembly of organismic bodies.

Both Stahl and Wolff's vitalistic theories were meant to account for the agentive, directedness of certain observed biological phenomena once their discoveries of mechanistic answers failed to do so. Whether it is Stahl's "active being" doing the driving inside organisms, or Wolff's supervisory agent playing the role of director, the vitalist hypothesis is one in which mechanism must be supplemented in some way with agentive properties in order to make sense of the observed organizational features (self-maintenance in Stahl's case, and growth, development and tissue differentiation in Wolff's).

[127] One difficulty with the notion of preformation is that it requires a large (possibly infinite) number of future generations to have their preformed selves nested in the current one like a Russian matryoshka doll, and yet no answer is given by preformation theorists as to how that structure gets constructed in the first place.

Élan Vital and Entelechy

The two most famous vitalist doctrines are also the most recent, each coming from the early twentieth century. The first of these is Henri Bergson's (1911) theory, laid out in his book *Creative Evolution*. Bergson was a philosopher, not a biologist; so we should note that his form of vitalism does not fit with the biologist's trend I've been describing. Still, it deserves mention primarily because his *élan vital*, or vital impulse, is probably the most widely recognized term for a vitalist substance today.

Bergson's primary motivation in *Creative Evolution* was the sense of conflict he felt between the continuity of life forms (each living cell has a parent, making every individual on earth part of one big continuous family) and the discontinuity of the same forms (species, for instance, are quite distinct, as are the differentiated tissues of a large multicellular organism). There are a number of parts to Bergson's theory but, regarding life's continuity in particular, he suggests that running through all organisms is a thread—his *élan vital*—which not only accounts for vitality but also for the passing down of that vitality through the generations, by way of the germ line.

The other famous vitalism is Hans Driesch's theory of "entelechies", a term he borrowed from Aristotle. Though it is somewhat debated what Aristotle meant by the term[128], Driesch borrowed it to refer to the force he suspected exists in living cells that directs their processes of development. Driesch was driven to this hypothesis when he separated the cells of a new two- or four-cell sea urchin embryo, and found that rather than growing into two half-organisms, as he

[128] "Entelechy" is usually translated as "actuality" in opposition to "potentiality", but just what that means and how it relates to other concepts is debated (Kosman 1969; Ross 1936; Charles 1994; Sachs 2005). Sachs (1995) notes that the word's roots in Aristotle's coinage are *entelēs* (ἐντελής, complete, full-grown) and *echein* (ἔχειν, to be a certain way by the continuing effort of holding on in that condition), but that, interestingly, it was also meant as a pun based in the terms *endelecheia* (ἐνδελέχεια, persistence) and *telos* (τέλος, completion). Sachs (2005) sums this all up by remarking that, on his interpretation, "Entelecheia means continuing in a state of completeness, or being at an end which is of such a nature that it is only possible to be there by means of the continual expenditure of the effort required to stay there." The resemblance of this definition to many of the components of the theory of teleology that will be offered later is rather wide.

suspected would happen, each grew into a complete individual organism (that is to say, his divided embryos grew into twins and quadruplets). The only explanation Driesch could muster at the time (this was half a century still before the discovery of the nature of DNA) was that the orderly division and growth of the cells was coordinated by an immaterial "mind-like" force (Driesch 1908). As with Stahl and Wolff, Driesch's experimental investigations brought to light some biological facts, but at the same time they highlighted deeper organizational questions about the principles of biology, which could only be answered in terms of an organizational tendency or force.

The Retreat of Vitalism

Clearly vitalism was a popular view. In fact, despite occasional mechanistic dissent (*e.g.* Descartes 1664/1985), it seems to have been the dominant theory of life from antiquity up through the middle of the nineteenth century, holding some sway even into the early twentieth century. By that time, though, its proponents simply ran out of steam and new patrons could no longer be enlisted.

The retreat was due largely to two factors. The first was the fact that while vital forces had often been hypothesized, an embarrassment of experiments had never once found evidence of them (Mayr, 2004, cites that there were literally thousands of experiments that came up empty-handed). The second factor was the accumulation of mechanistic advances that had been occurring in the golden era of mid-nineteenth-century biological theory—advances that seemed, to some, to finally explain or herald explanations for many of the phenomena that drove theorists to vital hypotheses in the first place.

To repeat our list, the phenomena that we are talking about are such vitalistic and organizational activities as embryological development, growth, self-maintenance, healing and regeneration, reproduction, spontaneous generation, animal movement, and heredity.

The greatest contribution to solving most of these riddles was the development and maturation of the cell theory, which was completed in the eighteen-fifties and consists of the discoveries that

(i) There are cells in biological organisms (Hooke 1665; van Leeuwenhoek)[129];

(ii) All living organisms and all their tissues are composed of cells (Dumortier 1832; Purkyně 1837; Schleiden 1838; and Schwann 1839)[130]; and

(iii) All growth and development occurs by the division and propagation of cells (Dumortier 1832; Remak 1852; Virchow 1855)[131].

Around the same time that the cell theory was coming to completion, a number of other important developments were also taking place in biology. Firstly, there was the organic chemistry revolution, which began with Wöhler's (1828) kidney-free synthesis of urea but was greatly accelerated later by the carbon-chain theory of organic compounds (Couper 1858; Kekulé 1858). What resulted was a realization that the compounds that make up cells and their products are the consequences only of

[129] Van Leeuwenhoek never published his work, but his extensive correspondence with the Royal Society in London has been preserved. A history of his discovery of microorganisms can be found in Dobell (1932).

[130] Schwann and Schleiden are usually credited with discovering the generalization that all living things are composed centrally of cells, but history is often blurry and imperfect. For one thing, Schleiden's work on plants duplicates Dumortier's earlier work and so it is unclear whether any credit at all should go to Schleiden. For another thing, the theory was not complete until animal and plant tissues were both seen to be the same, and so it is seems Dumortier (and perhaps Schleiden) deserve only partial credit for work contributing to the theory, since animal tissues were analyzed by Schwann and Purkyně. Lastly, while Purkyně (often written "Purkinje" in English) stated the generalization first, Schwann is usually credited for the claim because, for whatever historical reasons, his work happened to be more influential, or widely read. Perhaps all four should share the credit.

[131] Dumortier (1832) discovered cell division, or binary fission as it is also called; but it wasn't until twenty years later that Remak and Virchow were each able to combine that notion with embryological theories, such as Wolff's, in order to propose the generalization that growth occurs *only* through cell division.

chemical processes; the vital forces that had previously been hypothesized to be responsible for the construction of these larger molecules were unnecessary. Secondly, in the year following the carbon-chain theory, Darwin (1859) proposed his theory of evolution by way of "descent with modification", a further mechanization of the story of life that was able to account for the diversity and continuity of life forms. Thirdly, spontaneous generation was erased from the list of mysteries, when in 1862 Louis Pasteur showed it to be an illusion[132]. Fourthly, building on Lavoisier's and Laplace's (1780) work with guinea pigs that showed that breathing is essentially combustion[133], Helmholtz (1845) proved the conservation of energy in both physical and biological systems and thereby disproved the existence of the vitalistic notion of a nonphysical "animal heat". Fifthly, Emil Du Bois-Reymond, a colleague of Helmholtz's, showed with his (1848) physiology of nerve cells that animal movements were driven not by vital forces of any sort, but by electrochemical "action potentials". And, sixthly, over the next few decades, work on the nature of the chromosomes, found in the nuclei of cells, culminated in the discovery that these molecules—Schrödinger's "aperiodic crystals"—are somehow the vectors of heredity (Boveri 1904; Sutton 1902, 1903), although it took another half-century for the mystery of heritable information to ultimately be solved in detail (Watson and Crick 1953).

It really did seem as if the long list of mysteriously lively phenomena were ultimately being explained mechanistically, making any vitalistic hypotheses explanatorily redundant. But by the early nineteen hundreds, one vital question still had not been answered: If cells, their division, and their other physiological activities account for all the lively macroscopic mysteries that were being investigated . . . what accounts for the vitality of cells themselves? As before, with every proximal

[132] Pasteur duplicated and extended Spallanzani's earlier experiments with boiled broths (Vallery-Radot 1928). Life only appears in such Pasteurized broths when they are later allowed to be impregnated by air carrying dust particles that harbor cells. This evidence further cemented Remak's (1852) and Virchow's (1858) conclusions that cell division alone accounts for growth. As François-Vincent Raspail first put it, "omnis cellula e cellula"—all cells come from cells.
[133] Lavoisier and Laplace found that isolated guinea pigs produced the same amount of heat as chemical combustion that consumed the same amount of oxygen did.

mechanistic answer that biologists found, the question of what makes organisms lively was only pushed a little deeper, but never answered. Vitalism was dead, but the aspect of vitality had still not been explained.

Emergent Vitalism

There are two major interpretations we can make of the vitalist doctrine, and the notion sounds very different depending which of these interpretations we choose. I will call the two versions *enchanted* vitalism and *emergent* vitalism. The enchanted form of the vitalist conjecture is what we tend to see when looking back on historical writings. This is the thesis that there is something nonphysical in the world, something magical, spiritual, mystical, metaphysical, or what have you that somehow pervades living matter and brings it to life. The *élan vital* (or *entelechy, ens activum*, *vis essentialis, vis vitalis,* or *vis viva*) is a transcendental life-stuff. And this is the version that rightly offends the materialist. This is not to say that there *could not* be an *élan vital* (science has had much success with the hypothesis of invisible items that account for observations: space, atoms, electrons, fields), but it is only to say that all efforts so far have shown that there is no such stuff.

The emergent form of the conjecture is decidedly less extreme and posits only that there is some pattern—some still unknown, yet material, pattern—that is common to all living things and that accounts for their vitality. It suggests that life and vitality are made possible by something ultimately investigable that is consistent with physics and chemistry, and that very likely has to do with the *organization* of living things. This is, more or less, what was offered by the new breed of biological theorists, in the early twentieth-century, whose ideas came to be known as *organicism* or *holism* (Ritter 1919; Smuts 1926). In an attempt to reconcile the material and mechanistic constraints of the prevailing scientific worldview with the explanatory goals of vitalist thought, the organicists

developed and embraced the now commonplace concept that "an [organized] whole is more than the sum of its parts"—that is, that organization plays a crucial role in explaining certain phenomena, and from this tradition emerged the basis for the now more widely used scientific concept of emergence (see, *e.g.*, Goldstein 1939; Haldane 1931; Novikoff 1945; Ritter 1919; Ritter and Bailey 1928; Russell 1930; Smuts 1926; and see also more recent organicism-revivalist authors including, *e.g.*, Gilbert and Sarkar 2000; Kirschner *et al.* 2000).[134]

The organicists were right—organization does give rise to the emergent quality of vitality. All one needs to do to see this is to take some living thing and disrupt its organization just slightly in certain ways in order to see its vitality quickly disintegrate. For even a relatively large organism, something as organizationally modest as cardiac arrest, or the introduction of just a handful of the wrong type of molecules can result in terminal disorganization. Despite the power of their fundamental insight, the organicists lacked a more specific theory. There are plenty of things—shoe stores, hard drives, and even dead organisms—that are highly organized too, yet not vital in the least. What really needs to be sorted out (even still today) in order to account for the aspect of vitality is just what *kind* of organization it is that matters.

It didn't take long for the emergent vitalism of the early nineteenth century to fall out of favor. It was slow to produce fruit, weakened largely by a paucity of more specific hypotheses and, at the same time, the competing doctrine of materialism, which aimed to do away with all breeds of vitalism, was growing in power. As the mechanistic fields of molecular biology, genetics, and evolutionary theory came to be highly productive in the years leading up to the discovery of the structure of DNA and especially in the decades that followed, work on these topics eventually came

[134] The term "emergent" was coined for this usage by Lewes (1875). For a classic exposition of the generality and breadth of the concept in understanding what does and does not exist in the world, and how myriad phenomena can be more than the sum of their parts, see Holland (1998).

to dominate biological thinking. Concepts such as teleology and the holistic aspect of vitality were simply left behind in the wake of progress.

That is . . . until the nineteen-seventies, when a relatively small number of theorists resumed work in the emergent vitalist tradition by trying to sort out just what kind of organization it is that matters in bringing about life (Eigen 1971, Ganti 1971, Kauffman 1971, Maturana and Varela 1973). In Part II, we'll pick up the answer these theorists came up with, and develop it further. The general picture is that the kind of organization that matters in creating goal-directedness and the aspect of vitality is based in Kant's causal circularity, in which a thing is both cause and effect of itself. Before we can develop that theory, we need to be confident that we understand the shape of our subject properly, and so, first we must take up a mining expedition through another wide tract of teleological thinking that also established itself during the nineteen-seventies: the topic of function.

Chapter IV

Functions: Issues

Just as a physicist might say that heating a gas causes it to expand, a biologist might say that heating a mammal causes it to sweat. But a biologist might also say that a mammal sweats when heated in order to keep its temperature constant, while no physicist would say that a gas expands in order to keep its temperature constant—even though that is exactly what happens.

—David Hull (1974)

What, then, are the theoretical commitments implicit in the biological concept of function that distinguish the case of sweating from that of the expanding gas? Why is constant temperature merely an effect of gas expansion while being the 'function' of sweating in mammals?

—David Buller (1999)

By far the bulk of recent literature in the field of teleology has concentrated its attention on answering the question, "What is a function?"[135] As Aristotle put it through his use of the concept "that for the sake of which", there is an undeniable temptation to see a relationship between function and purpose. Whenever we ask what the function of a particular item is, it seems we could just as well replace the question and ask what the item is for, or what its purpose is.

For example, if we ask, "What is the function of the standing rigging on a sailboat?" a reasonable answer would be that it is to stabilize the mast and to transfer force from the sails onto

[135] Though the trend has accelerated since about the mid nineteen-sixties, the focus has been on functions for roughly eight or nine decades now, ever since vitalism lost its appeal.

the hull. Not only is the rigging *able* to stabilize the mast and transfer the wind's force, but that is very clearly what it is *for*; it is what it is there for and it is what it is meant for, it is what it was designed for, intended for, made for, and used for, and it is something it is usually particularly good for. There is no doubt that that is what we would say is its purpose. The same could be said of a biological trait such as the heart. We can ask, "What is the purpose of the heart?" and get the same answer as when we ask, "What is the function of the heart?" Our sense is that the heart is *for* pumping blood. It is good for that, made for that, used for that, and it is there for that. We (modern materialists) understand the heart's being there in terms of a non-human sense of design—in terms of the natural design of evolution by natural selection—but nonetheless we understand pumping blood as the heart's purpose and as its function. It really seems as if the questions "What is it for?", "What is its purpose?", and "What is its function?" are interchangeable since, in many cases, the same explanations answer all of them. In this sense, having a function appears to be a centrally teleological phenomenon.

The issue, however, is not nearly so clear-cut as this. Indeed, many authors find the notion of teleology completely unacceptable because of concerns such as backwards causation and the subjective, value-laden nature of it. These writers claim, despite the kind of obvious intuition above, that biologically functional items could not be *for* anything, since natural selection is a mindless rather than a foresighted process, a series of happy accidents without anyone behind the wheel who might intend or mean or cause the items of biology to serve their design objectives (*e.g.*, Cummins 1975, 2002; Davies 2001). The position these authors take is that our tendency to describe biological items as "being for" something is merely a result of imaginative reflection . . . overactive analogizing, because teleology, they claim, is an illusion.

Debate over this topic is lively. I'll introduce the numerous camps later, but for now it is enough to say that there are still other authors who, while seldom claiming that functions are

teleological, will nonetheless contend that we can understand many things as being "there for" something or as having been "designed for" something.

My own position, which I'll begin to lay out in this chapter but which won't be fully fleshed out until the end of Part II, is that function*ing* is based on an evaluative kind of norm that justifies both the use of the word "for" and our understanding of function as a purposive notion. One thing that sets my view apart is that it is neither anti-teleological nor grounded in the concepts of being "there for" or "designed for". As we will see, my view is also set apart by not accepting the notion of *a* function as being a pattern or phenomenon in the world centrally worth analysis. I will argue that *function* is an illusion—a useful fiction much like colors—and that *purpose* or *goal-directedness* is the more fundamental pattern to be analyzed.

Chapter Guide

Still, since a theory of teleology would be remiss if it did not address the extensive existing literature on functions and if it also did not somehow account for our intuitions about the functioning of things that seem to have functions, I will spend this chapter and the next looking at the phenomenon of function, the concept that reflects it, the intuitions that accompany it, and the theories that have been presented to explain it.

A significant portion of the analysis in these two chapters will be an unabashed conceptual analysis, a method that I defended earlier (pp. 29–39). To recapitulate: I am using conceptual analysis not to develop a theory, but rather to outline a subject and to highlight the points in potential theories where significant inconsistencies with either intuitions or facts exist. Here, I would like to add one more justification for my use of the method: the bulk of the arguments presented in the existing literature on functions is itself based substantially on intuition and so,

before offering an alternative view (in Part II of the dissertation), I feel it is necessary at least to neutralize if not to reverse the appeal of some of those perceptions in favor of new ones. I will attempt to reshape the way we perceive the subject before I offer an alternative theory of that subject. As I mentioned above, unlike the majority of theorists in this field, I don't believe that functions exist, yet I do believe that goal-directedness does, and those differences in belief already produce a radically different subject about which to theorize. In addition, I believe the focus upon function instead of goal-directedness is supported by a broad conceptual edifice of problematic (yet very intuitive!) instincts that has significantly hampered progress in the field; in order to move forward, this entire set of ways of thinking must be overcome by carefully reconsidering many of our basic assumptions about the topic. A large part of what is to come in the next two chapters, then, will consist in questioning the intuitions of previous theorists and presenting new alternatives by using, among other things, the tools of (cautious) conceptual analysis. It will be the job of Part II to provide a theory that can make sense of the new blend of intuitions that result from this analysis.

With the foregoing considerations in mind, the rest of the chapter is laid out accordingly: First I will discuss the autonomy of biology among the sciences and the role that some think is played in it by the notion of function. Then I will introduce the concept of function in broad strokes, mentioning some of the major observations that have been made about it. Following that, we'll look at how we use that concept in statements we make about the world, which comprise one form of "data" to which a definition or theory of the pattern underlying that concept might (limitedly) be held accountable. Then I'll try to develop some intuitions about what we mean when we say an item *has* a function and how this differs from an item *serving* a function. There, I will begin to present my argument that items in the world do not in fact have functions. In the two sections that follow, I'll review two common intuitions that I believe are highly misleading and that tend to confuse much of our thinking about function. The first is the intuition that the phenomenon of

design entails function. The second is one of the intuitions discussed above—the idea that "What is it for?" is equivalent to "What is its function?" Lastly, I'll review the concept of *accidents*, which has long been used to evaluate theories of function, and I will ultimately show it to be an unproductive strategy—a red herring.

A. The Autonomy of Biology

No evidence has been unearthed in our inquiries into genetics and molecular biology that would argue positively and persuasively for the inherent autonomy of biology. Moreover, since genetics occupies a central position with respect to the problem of growth and differentiation of an organism, there is evidence that these processes will eventually admit of a complete chemical explanation.

—Kenneth Schaffner (1967)

Functional explanation is the trademark of biology. Whenever it is necessary to assert the autonomy of biology against a reductionistic threat, the important role of functional explanation for a proper understanding of biological phenomena always features prominently in the argument.

—Manfred Laubichler (1999)

Schaffner and the other reductionists that Laubichler is concerned about in the epigraphs above might challenge us with a claim something along the lines of "Well, biology is just biochemistry—a species of chemistry—but the fundamental rules remain unchanged from those of any other chemistry: thermodynamic imperatives still hold, mass and energy are still conserved, reactions still proceed only as allowed by energies of activation and catalysis, and so on . . ." Such a reductionist view is driven by the reasonable belief that there is no magic in biology—no extra-physical substance or process that generates the liveliness in life. The *entelechy* and *élan vital* of Driesch and Bergson's respective metaphysics were never found, and vitalism died a cold death in the last century.

While we wouldn't want to deny the details of the reductionist's claim about the contribution of biochemistry to biology, some, such as Laubichler, observing what Monod (1972) called the projective behavior of biological organisms, may want to question the use of the word "just". Does biology *just* follow the rules of biochemistry?[136] Or is there something more to being lively and projective?

The Emergent Perspective

Ernst Mayr, in his assertion of "The Autonomy of Biology" is one who wants to draw our attention to the possibility of taking the emergent perspective on biology.

> Nothing is as characteristic of biological processes as interactions at all levels, among genes of the genotype, between genes and tissues, between cells and other components of the organism, between the organism and its inanimate environment, and between different organisms. It is precisely this interaction of parts that gives nature as a whole, or the ecosystem, or the social group, or the organs of a single organism, its most pronounced characteristics. To repeat what I said before, rejecting the philosophy of reductionism is not an attack on analysis. No complex system can be understood except through careful analysis. However, the interactions of the components must be considered as much as the properties of the isolated components. And this is what the reductionists had neglected (Mayr 2002).

[136] In the same paper cited above, Schaffner continues his reductionist disparagement of the autonomy of biology: "The antireductionist biologist, accordingly, seems to be restricted to asserting a type of "make-believe" autonomy. He may plan, execute, and interpret his experiments without worrying about reduction to a molecular level, but this is no reason for maintaining that a biological entity is anything more than something ultimately characterizable and explicable by molecular biology. [. . .] This make-believe autonomy may well be heuristically valuable, though perhaps relative to a particular stage of development of the sciences. There seems to be no positive evidence, either logical or empirical, for any real autonomy." (Schaffner 1967).

Perhaps what is special about living things and their unusual, projective behavior, Mayr reminds us, is not the parts that they are made of, but something emergent, something about their organization. But, of course, as Mayr's list points out, presuming so leaves us mired in complication by the numerous levels and kinds of interactions that occur in biological systems. Biology, like any major branch of science, is not a unitary phenomenon. The question, then, as to whether biology as a whole can be reduced to chemistry and physics may in fact not be well-formed. Instead, we might be better off asking individually whether the concepts of ecology can be reduced, whether the concepts of morphology can be reduced, whether the concepts of natural selection can be reduced (see Rosenberg 2006), whether the concepts of genetics can be reduced (see, *e.g.*, Waters 1990), and whether the concepts of cellular processes and those of molecular biology can be reduced (see Campaner 2010; Elgin 2010; Sarkar 1998; Schaffner 1967; Weber 2005).

There is no consensus on any of these questions yet, but as our understanding of many of these subfields progresses, it is becoming increasingly evident that many of the concepts in many of these subfields may largely, if not entirely, reduce. Even if scientists haven't yet worked out the minutiae of each, the general modes by which such reductions would take place are becoming increasingly well understood. One question that remains for the emergentist is, once the molecular details of biology have been worked out in full, whether or not there might be any organizational principles of biology that remain irreducible. Are there any candidates for general and *thoroughly biological* concepts that evade reduction and that thus could be considered emergent, purely biological, laws?[137]

[137] The term "law of nature", interpreted to mean completely exceptionless and non-derivative generalizations about the universe, is itself a controversial notion. Some philosophers don't think such laws exist, even in physics. Others imagine they may be unknowable. And others think maybe they exist but they are as yet unknown. It is often argued, for instance, that Newton's law of gravitation was ultimately found not to be exceptionless; why then should we assume any of our currently accepted laws to be? (*e.g.* Cartwright 1980). For our purposes here, though, we can somewhat relax the

Natural Selection As a Law?

The most obvious contender might be natural selection. It is certainly relevant to just about everything we consider biological and, if we take it to be, in Dennett's (1995) terminology, a rote algorithm—a kind of mathematical or logical principle that is entirely substrate-independent—then it seems very law-like in its generality. A number of authors have argued that it is indeed a fundamental law (see *e.g.* Dawkins 1983; Eigen 1983; Rosenberg 2001a, b, c, 2006). Rosenberg (2006) points out that natural selection can occur on objects that he considers to be non-biological, such as autocatalytic, self-replicating molecules, and so he classifies natural selection as a *physical* rather than biological law, but a law nonetheless (see also Eigen 1983). The argument is compelling, but there is also a second way to interpret this point—we could simply consider such replicating molecules to be "alive" or "biological", thereby keeping natural selection in the realm of biology. That suggestion might offend the sensibilities of some theorists but not of others, though I suspect that is only because what is meant by the terms "alive" or "life" is not yet agreed upon. For the moment, there is no need to resolve this issue; we will look at it again near the end of the dissertation.

Function as a Law?

Another possible contender is given by Laubichler's reply to the reductionist: biological objects *have functions* (see also Ayala 1968; von Bertalanffy 1968; Mayr 1974/1988, 1992, 2002; and Williams 1966 for similar ideas). This answer is a good candidate, too, because function seems to be a general, biologically relevant, organizational concept far removed from the vast variability that

---

definition of "law": We are just interested in knowing if there are any general, predictive, and explanatory principles that underlie phenomena in biology.

exists in lower level biochemistry, cell biology, genetics, and ecology. As we'll see, though, we don't currently have any theory of how function might reduce to physics and chemistry, and some even argue that it fundamentally cannot. The main trouble, however, with attempting to develop a theory of function is that biologists and biophilosophers have been unable to agree upon precisely what the term "function" means and, without a definition, we have no basis upon which to either argue for or attempt to disprove reducibility. For now, Laubichler's answer remains somewhat hollow, as it only replaces one undefined term ("life" or "biology") with another ("function").

Multiple Realizability

Some writers, particularly in philosophy of mind, have argued, in terms separate from biology, that functional kinds (such as *wing*, *eye*, *handle*, or *brake*) are in fact irreducible to physics. The argument, which goes by the name of "the multiple-realizability thesis" (see, *e.g.*, Block and Fodor 1972; Fodor 1974; Kim 1996; Putnam 1975; Turing 1950) contends that since functional kinds can be physically realized in various ways and still perform "the same" function—that is, for instance, since either glass or Plexiglas or borosilicate or sapphire crystal can fulfill the function of letting light through (as in a window or lens)—there cannot be any kind of general, natural law defining functions in terms of their physical constituents or processes. According to the argument, the best generalizations that can be made about functional kinds will describe physically heterogeneous (disjunctive) categories (such as the category "glass *or* Plexiglas *or* borosilicate *or* sapphire *or* . . .").

Although the argument is at first blush quite convincing, there is some debate about whether Putnam, Block, Fodor, and Kim are right about this (see, *e.g.*, Churchland 1986; Lewis 1969; Richardson 1979; Weber 2005). For my part, I think that the multiple-realizability thesis may be

cause for concern about reducing individual functional kinds to general physical categories, but that it would miss the point if it were applied to the more general category of *being functional.* I think "being functional" is a general pattern that, while accounting for quite heterogeneous types of effects (flight, sight, lifting, braking, and so on . . .) is likely to be realized by either a monopoly or a small oligopoly of structures in the same way that *chemical bonds* account for an exceedingly diverse array of chemicals (N_2, H_2O, CO_2, NaCl, $Fe_2O_3 \cdot nH_2O$, and on and on . . .), despite coming in only a few structural variants (ionic, covalent, polar covalent, dative, metallic, and hydrogen).

Multiple Laws?

It seems to me that there are three astounding high-level facts about biology. The first is that there are things that are alive. The second is that there is such an amazing diversity of things that are alive. The third is that the many and diverse things that are alive are also so well adapted to their environments. Perhaps three distinct (yet possibly related) explanations can account for these three facts.

McShea and Brandon (2010) have suggested that "Biology's First Law" is "the tendency for diversity and complexity to increase in evolutionary systems". Their argument is that this tendency is a statistical result of random change in reproducing populations, *regardless of whether natural selection occurs.* A random walk through the space of possible organisms will behave the same way any random diffusive system does—eventually, the diffusing agents will explore all of the space excepting regions where there are barriers to exploration. Diversity is a likely result. I am inclined to agree with McShea and Brandon that the simple mathematical generality of their proposition also

warrants law-like status *and* that it precedes natural selection[138]. However, since their idea depends on the prior existence of living things, it appears to me that perhaps this concept deserves to be called biology's *second* law.

I agree also with Rosenberg's verdict: Natural selection appears to be an organizational principle general enough to be called a law, and while its applicability seems to generalize beyond genetic cell-biological systems, it appears to be largely explanatory of much of the realm of biological phenomena. In particular, it explains impeccably the tendency of biological items to quite commonly be adapted.

From the foregoing discussion, two questions remain. First: Is there another law of biology, logically prior to McShea and Brandon's diffusionary principle and Darwin's (1859) theory of natural selection, able to account for the vitality of living things? And, second: What do we make of Laubichler's (and others') proposition that functional explanation accounts for the autonomy of biology? I think the answers to these two questions will be convergent, though the answer to the second will turn the question on its head. That is to say, by the time we sort out what it means for something to function, we will also understand what it means for something to be alive; it won't, however, be *function* that brings about vitality but vice versa (after all, cogs and levers are functional but not in any way lively, on their own).

[138] While some might argue that the tendency to diversity has always been an assumption of the theory of evolution by natural selection, I agree with McShea and Brandon that it can be conceptually useful to distinguish various components.

B. The Concept of Function

It is necessary to conclude that the blood in animals is driven around in a particular circular motion; and that it moves perpetually, and that this is the action or the function of the heart, which by pulsation it performs, and it is entirely the motion and the pulsation of the heart that is the only cause.[139]

—William Harvey (1628)

What exactly do we mean when we ascribe a function to an object—what property or properties are we attributing to that object by such an ascription? That is, when we say, for instance, that pumping blood is the function of the heart, what is it about pumping that makes *that*, as opposed to anything else the heart does, its function? Put another way, when we inquire as to what the function of some object (such as the heart) is, what precisely is it that we are asking for? What kind of an answer will satisfy us? This is the sort of question that lies at the heart of defining the concept of function.

The answer to an inquiry about *a* function—that is, an *individual* function (say, of the heart) rather than about function(ing) itself—usually takes the general form of a description of something that the item does in some manner, under some circumstances: "That particular thing the heart does that _____." But there is some debate as to just what phrase should fill in the blank here. The following list paraphrases some of the more prominent attempts; in the next chapter, we'll look at each of these in more detail.

[139] This quote has been translated. The Latin original reads: "Necessarium est concludere circulari quodam motu in circuitu agitari in animalibus sanguinem; et esse in perpetuo motu, et hanc esse actionem sive functionem cordis, quam pulsu peragit, et omnino motus et pulsus cordis causam unam esse." (William Harvey 1628).

Causal Roles (CR): That particular thing the heart does that explains its role in the circulatory system that it is a part of (Cummins 1975, 2002; Davies 2001).

Selected Effects (SE): That particular thing the heart does (or, rather, that its ancestors did[140]) that accounts for why the heart has come to be there (Ayala 1970; Buller 1999; Griffiths 1993; Godfrey-Smith 1994; Neander 1983; Millikan 1984; Wright 1973).

Replication Dispositions (RD): That particular thing the heart does that explains how it contributes to the potential creation of future copies of itself (Bigelow and Pargetter 1987; Canfield 1964; Griffiths 2009; Lewens 2004; Ruse 1971).

Goal Contributions (GC): That particular thing the heart does that contributes, in a general sense, to the achievement of a goal of some sort (Boorse 1976, 2002).

Programmed Effects (PE): That particular thing the heart does that it has been programmed to do (Mayr 1976/1988, 1992, 2002).[141]

Valuable Effects (VE): That particular thing the heart does that is good for its bearer (Bedau 1991, 1992, 1996, van Parijs 1982).[142]

---

[140] The term "the heart" can be ambiguously interpreted as either a *token* descriptor—an individual heart—or a *type* descriptor—the category of all hearts. The selected-effects account is implicitly a type analysis, so it is better to think of the term, here, as referring to a class of hearts—the *reproductively established family*, in Millikan's (1984) terms—that a particular token may be a member of.

[141] Mayr's analysis does not directly address the concept of function. However, he makes it clear that this is because "the word *function* refers to two very different sets of phenomena" wherein it is "sometimes used for a physiological process and sometimes for the biological role of a feature in the life cycle of the organism" (see also Bock and von Wahlert 1969). Mayr finds biological roles to be teleological, but "physiological functioning of an organ" not to be (cf. the pluralist view described soon). The quibble is minor: What Mayr calls a function is not what most others do, but his theory of the teleology in biological roles is comparable to others' theories of functions (1992, p. 123–4).

Not only is there debate as to which phrase might fill in the blank to complete the answer, but there has also been some doubt expressed about the idea that only one such phrase might do. One of the more common contemporary views is what's known as the pluralist account and, far and away, the most common pluralist account is a hybrid that relies on both the causal role and the selected effects theories, employing each to explain some but not all uses of the term "function" (see, for example, Millikan 1989b; Godfrey-Smith 1993; Allen and Bekoff 1995).

Perhaps the easiest way to understand pluralism is to note that different senses of the word "function" mean different things. The pluralist suggestion is that there may simply be more than one core concept that we refer to as a function, and so we need multiple, parallel (disjunctive) definitions in order to account for all the ways we use the word. As with any word, there certainly are multiple senses to the term "function" and we will explore many of them in the next section; however, I think the central phenomena in which all the theorists above are interested are actually unitary, and so, rather than turning to pluralism, I will give an account that attempts to unify them.

---

[142] While van Parijs's account is directly aimed at the concept of a *function*, Bedau avoids the term and claims that his is an account of *teleology*. However, Bedau casts his discussion as an analysis of "in order to" statements, which are generally viewed in the literature as function statements (see Beckner 1969, *e.g.*), and he compares it directly with other function analyses, thereby adding it to that class implicitly. At the least, then, we should agree with Bedau that his theory of teleology can be compared with function theories to the extent that those other theories are teleological.

C. Function statements

The function of chlorophyll in plants is to enable plants to perform photosynthesis (that is, to form starch from carbon dioxide and water in the presence of sunlight).

—Ernest Nagel (1961:403)

The function of gills in fishes is respiration.

—Francisco Ayala (1968:218)

It is commonly taken that one central requirement for a theory of function is that the theory be able to make correct attributions of functions—to say with some authority that some feature is or is not a function of an item—and so theorists typically measure the success of their claims against a set of function statements that we would intuitively agree to be true. If a theory cannot straightforwardly declare that the function of the heart is to pump blood, then it seems that either we must find a way to understand the heart's pumping as a non-central example (or non-example) of what the theory purports to account for, or else something about the theory needs to be revised.

Some philosophers of science (*e.g.*, Neander 1991; Wouters 2005) find this to be too restrictive and would suggest that a theory of function need only account for the statements made by *scientists* in their practice, since there may be differences between the concept as it is used by specialists and as it is used by the rest of us[143]. Others who find (conventional) conceptual analysis

[143] However, see McLaughlin (2001:86) for an argument that "function" is not a technical term—it is not yet well-defined—and so, for one thing, what we mean by "function" and what biologists mean by "function" might differ little, if at all and, for another thing, biologists' concept of function may not be justifiably authoritative.

to be distasteful (*e.g.*, Millikan 1989a) refuse to be held accountable to our intuitions about function statements at all.[144]

I agree that it is possible, even likely, that our ordinary concept may not align with a specialist's concept, much less with the structure of the world. Nonetheless, I find function statements to be instructive. Even though they cannot serve to produce a theory in the conventional conceptual analyst's sense, function statements will be useful to us in grouping many of the phenomena in which we are interested, in revealing some of the regularities we perceive therein, and in serving as a preliminary form of data to ensure a certain kind of internal coherency of our theory. The more central examples can be used to make false theories seem unlikely, and the more peripheral examples can be used to help us understand the fringes of a theoretical offering. Also, as we'll see below, analyzing function statements themselves (and categories thereof, and regularities therein) can help to produce some insights that may clarify the subject we will be theorizing about.

The Senses of "Function"

As was noted in the previous section, "function", like most words, is used in a number of varying senses and so it is important for us to spell out just what sense or senses of the word interest us, in order to understand which function statements should guide our theorizing and which others might simply be distracting. I will largely follow other theorists in developing these sense-categories but, as will be seen, my interpretation of the key sense will differ significantly. Even so, I think the

[144] While Millikan takes her account to be a theoretical definition, not a conceptual analysis, she still claims that her "definition of 'proper function' may be read, roughly, as a theoretical definition of function . . . in the context 'The/a function of _____ is _____'." (1989a:291). This claim groups a certain set of implicit function statements that agree on a certain sense of the word. And whether or not she claims her project to be conceptual analysis, her theoretical definition may be held accountable to that implicit set of function statements, for if the definition didn't make sense of central cases such as "the function of the heart is pumping blood" then no one would take her definition seriously.

theory I offer later in some way accounts for all the function statements that previous theorists have concerned themselves with, but it does so by providing different accounts of each of the senses and so it is useful to individuate those senses first.

Wright offers eight examples of the word "function" as used in sentences, showing some of the range of meanings it can convey, and he uses this list to zero in on the sense of the word that he thinks is under scrutiny in the project of defining biological function. So far as I know, no one has disagreed with his assessment, at least in print. Wimsatt (1972), following Nagel (1961), produced a similar list of word-senses, the details of which we'll also look at shortly. Here are Wright's examples.

W1. $y = f(x)$ [or] The pressure of a gas is a function of its temperature.

W2. The Apollonaut's banquet was a major state function.

W3. I simply can't function when I've got a cold.

W4. The heart functions in this way . . . (something about serial muscular contractions).

W5. The function of the heart is pumping blood.

W6. The function of the sweep-second hand on a watch is to make seconds easier to read.

W7. Letting in light is one function of the windows of a house.

W8. The wood box next to the fireplace currently functions as a dog's sleeping quarters. (Wright 1973)

To the list we could add a handful of similar ones:

W9. The computer science instructor asked the class to write a "quicksort" function as homework.

W10. The wristwatch performs four functions in addition to telling time.

W11. In their function as legal adviser to the President, the Attorney General will give their opinion on matters in which the President is an interested party.

W12. The bible in the soldier's breast pocket functioned to stop a bullet from entering his heart (an example of this sort is used by Wright, 1973; and again by Boorse, 1976, 2002).

Some of these can be discarded quickly. The meaning in W1 is derived from the idea of a relationship or process that maps items from one category or set (x-values or temperature values) onto those in another (y-values or pressure values). In this sense, the relationship or process is derivatively named "a function" in order to indicate that it is an entity that serves a particular (mapping) purpose. Leibniz appears to have been the first to extend the word this way; in his (1673) *Methodus Tangentium Inversa, Seu de Functionibus* he uses the term analogically to refer to the idea that a mathematical relationship performs a certain job (Swetz *et al.* 1995). The meaning in W9 is further extended from the Leibnizian sense—functions written in computer languages are essentially instructions that carry out mathematical operations of varying complexity; functions of this sort are also more clearly conceived of as pieces of code that may be executed in order to accomplish a particular job (for a particular purpose). Similarly, the meaning in W2 derives from the idea of formal gatherings serving certain (social) functions or purposes. Wright suggests that both word senses W1 and W2 (to which I have also added W9) are irrelevant to the central question of function

that philosophers of biology and teleology are interested in (Wimsatt, 1972, makes a similar judgment). I am happy to take that recommendation, and we'll see why in just a few moments.

Wright then names examples W5, W6, and W7 as the paradigm cases. These are said to be the kinds of statements or claims that we should be interested in when we ask, "What is a function?" They are the functions *of* objects, parts, and traits. In contrast, he points out that W3, W4, and W8 (to which I will add W12) seem to be peripheral cases. The best way to make this distinction straightforward is to recognize these latter examples as cases of *verb-function* and examples W5, W6, and W7 as cases of *noun-function.*

The reason why W1, W2, and W9 were discarded earlier is a little clearer now if we continue this grammatical-role analysis: while each is also a type of noun-function, it is an *object* noun that, itself, refers to some entity, whereas the noun-functions in W5, W6, and W7 are all *properties*—the latter are not objects but things *had* by or *served* by other objects or entities. In order to simplify this important distinction, we can call W1, W2, and W9 "object-functions" and W5, W6, and W7 "property-functions" or, as is the custom in the literature, "proper functions" (Millikan 1984; Neander 1983). The kind of functions that traditional function theorists seem to be centrally interested in are not just noun-functions but, more specifically, these proper functions.

A possible question arises, though: Does verb-functioning derive its meaning from the existence of proper functions or vice-versa? That is, does having a function, in the first place, allow a thing to function in a certain way . . . or does functioning in a certain way allow us to say that a thing has a function? In the next section, and throughout the dissertation, I'll defend the idea that most theorists' ultimate focus upon proper functions is misplaced. I will argue that proper functions depend intimately upon verb-functioning and that our theory should reflect that, and that the very idea of *having* a function is in fact an illusion. Nonetheless, I take it that a good theory of verb-functioning should also explain clearly just how that illusion of proper functions arises from the

existence of verb-functioning, in order to justify—or at least explain—both the scientist's and the layperson's usage in attributing proper functions.

We can look at the remaining examples now: under this grammatical-role rubric, examples W10 and W11 should be categorized along with W5, W6, and W7. The many functions of a modern wristwatch or pocket calculator or smartphone are all instances of proper functions. Likewise, for a person to "have" a function (in their job), means they are meant to serve in a particular role for a particular purpose—the function is a property of the person or of the role the person is filling.

Fringe Cases

Some fringe cases are worth pointing out now. Lewens notes, "A physical chemist may ask 'What is the function of free radicals in the breakdown of atmospheric ozone?'" (2004, p. 119). This usage corresponds closely to another use of the word: In organic chemistry, scientists refer to a subset of the atoms that make up a molecule as a functional group or sometimes just as a function. "A functional group" or "a function" may at first appear to be an object-function, but what the name highlights is the characteristic set of reactions of the class of compounds that share this group of atoms, and so one might think of the propensity to take part in those reactions as being the property-function of that group of atoms, though the case is an outlier precisely because, while it is a property-function, traditional theorists wouldn't consider it to be a proper function (the same goes for the function of free radicals in the breakdown of atmospheric ozone). Lewens' example suggests another similar one: it can also be asked, "What is the function of the ozone layer in our atmosphere?" to which a coherent answer, involving the filtering of radiation or possibly the disintegration of meteoroids and other space debris, could be offered. Similarly, we would accept an answer involving nitrogen fixation, ecosystem-level nutrient cycling, or vegetative spread to the

question, "What is the function of rhizomes in the forest ecosystem?" (see *e.g.* Guo *et al.* 2008; Helmisaari *et al.* 2007; Iversen *et al.* 2015). These two examples appear also to be property-functions yet are generally not considered to be proper functions.

For now, to be reasonably inclusive, we will retain all these examples of properties that are called functions on our list below, but we should note that they do differ somewhat from Wright's paradigm cases. We can follow Lewens, who suggested that the kind of claim in the free-radicals example will rarely be used "in concert with terms like 'purpose,' 'problem,' 'solution,' or 'design'" (2004, p. 119). The same holds for functional groups and the function of the ozone layer in our atmosphere. These are nonetheless interesting fringe cases both worth consideration and requiring explanation of one sort or another. I think the use of an analogy in our thinking explains why the word "function" applies to them, but describing the source of the analogy that allows the word to be further extended in this way will have to wait until we have a full theory of the more central cases.

A Summary of Senses

To summarize, we have analyzed the concept of function into its usage in four main categories of function statements. Though it may be possible to further subdivide the categories, I think this level of granularity will serve us for what we need.

[1] Proper functions, which have historically served as the central focus of the literature on function, are used to describe the functions an item has. "The function of the heart is to pump blood", "a function of a window is to let in light." I claim that these functions derive their status from verb-functions, though the manner in which they do so will require further argument.

[2] Verb functions are used to describe a role that an item plays. "The heart functions *as* a pump . . . ", "the box functions *as* a dog's sleeping quarters", and "the bible functioned *to* stop a bullet." This sense of the word "function" will be of central interest for our later theory.

[3] Fringe functions are a kind of property-function that derive analogically from verb-functions, in a manner different from how proper functions do. No theorist has yet seriously looked at or made provisions for this category, for good reason, but nonetheless I think we can produce a satisfactory explanation for its existence and I will do so later.

[4] Object functions are the least interesting category for our analysis. These are objects called "functions" that have derived their names, by historical custom, from the particular property-functions that they serve. For instance, mathematical functions, computer programming functions, and events as a kind of function.

The central sense in most analyses is proper functions but I will argue that this sense is only properly understood in terms of verb-functions and so, eventually, a fully explanatory theory will have to offer analyses of both. In the meantime, here is a list of (both proper and fringe as well as verb) functions that I will refer back to, later in the dissertation.

FS1. The principal function of a mammalian heart is as a pump to circulate blood through the vasculature of the body.

FS2. In many birds, the function of wings is to allow flight.

FS3. In many birds, one function of feathers is to aid in flight; another is to serve as insulation; a third is to act as waterproofing.

FS4. The function of the sweep-second hand on a watch is to make seconds easier to read.

FS5. Letting in light is one function of the windows of a house.

FS6. The function of a cog or a lever, in a machine, is the transference of rotational mechanical force.

FS7. The function of a rock, used for some time as a paperweight, is to hold down papers on a desk.

FS8. Stabilizing the mast and transferring the force from the sails onto the hull are the two main functions of standing rigging on a sailboat.

FS9. "The function of a telephone is effecting rapid, convenient communication." (Wright 1973)

FS10. The function of putting a stamp on a letter is to ensure that the letter gets delivered.

FS11. The function of writing to one's congressional representative is to influence their decision.

FS12. In their function as legal adviser to the President, the Attorney General will give their opinion on matters in which the President is an interested party.

FS13. "The wood box next to the fireplace currently functions as a dog's sleeping quarters." (Wright 1973)

FS14. The bible in the soldier's breast pocket functioned to stop a bullet from entering his heart. (Boorse 2002; Wright 1973)

FS15. Protecting the long-term interests of the bourgeoisie is the function of the capitalist state (van Parijs 1983).

FS16. Maintaining social cohesion is the function of religious ritual (van Parijs 1983).[145]

FS17. "What is the function of free radicals in the breakdown of atmospheric ozone?" (Lewens 2004)

FS18. One function of the ozone layer in our atmosphere is to filter harmful radiation out, preventing it from reaching the surface of the planet.

FS19. The function of rhizomes in the forest ecosystem is the fixation of nitrogen.

FS20. The function of a soliloquy in a play is to inform the audience of a character's inner ruminations.

One thing that emerges quickly from this list is that objects or traits do not necessarily have or serve just one function. The functions of feathers (FS3) are a clear example from biology, but multitasking happens in engineered artifacts as well: in a diesel engine, the injectors have a return line through which excess fuel flows back to the tank during the injection cycle. The return of the excess fuel is one function of this backflow; two more functions are the cooling and the lubrication of the injector mechanism by the fuel as it flows through it. Currently, all modern theories of function can be made consistent with the fact that functions of objects are not unitary, so the issue of multi-functionality is not a major sticking point, but we need to ensure that any new theory also respects this fact.

Another thing that emerges is that there seem to be just a few major classes of items that have functions—organisms and artifacts (though perhaps more properly, *traits* of organisms, and artifacts and their *parts*) and behaviors (such as FS10 and FS11). The first two are commonly observed, while the third is only occasionally so. One might also argue that social structures

[145] Van Parijs' examples are included because they come from a non-central field (neither organism nor artifact) and so make an interesting case. They are both clearly property-functions.

(institutions) and behaviors make for different categories than organismic structures (traits) and behaviors, in which case there are five main categories. Aside from van Parijs (1981), most recent writers have neglected social functions in favor of biological and artifactual functions; I include examples FS15 and FS16 for the sake of generality. There also seem to be some exceptions to these three (or five) categories. For instance, neither the ozone layer nor the free radicals in it (FS17) are artifacts, organisms, traits, behaviors, or social structures, and a rock simply chosen as a paperweight (FS7) seems to blur the boundary between artifact and mere object. Our later theory will need to clarify both the reason why functions seem to apply primarily to artifacts, organisms, and behaviors, and the reasons why exceptions such as these appear to occur.

The Perspectival, Evaluative, and Teleological Senses

Under Wimsatt's (1972) analysis, our concept of proper function divides into three more specific senses. He eventually suggests that all three may be roughly the same, and I agree. But it is worth noticing the differences, at least at first, to get a sense for the ways that the concept can be stretched, and some of the common features that have been noticed.

Using the *perspectival* sense of "function" means that only when viewed from a certain perspective are some of the causal consequences of an item its functions. However, there is not necessarily a single privileged perspective from which such analyses can be made. Wimsatt borrows this inherently subjective, context-sensitive sense from Kauffman's (1971a) discussion of "parts explanations", though it is also closely related to the CR analysis, which I'll discuss more in the next chapter. For Kauffman, it is important to note that there are various overlapping ways to divide items into parts and each provides different possible perspectives that highlight different functional roles that may be played.

Wimsatt's *evaluative* sense of "function" derives from Lehman's (1965) and Hempel's (1965) notions in which a function exists when a part contributes to the proper operation of a system.[146] Proper operation, obviously, requires a standard of evaluation, and so evaluative functions are relativized to such a standard. It is an open question as to what could provide this standard, but assuming for now that the intentions of a designer or a user might provide a standard, then in this sense a light bulb has a function only when it works to produce light. This is at odds with Millikan's intuition (which we'll review soon) that a malfunctioning item still has a function. The valuable-effects analysis of function (that we'll see in Chapter V) is an attempt to turn this sense into a theoretical claim.

Wimsatt's *teleological* sense of "function" means that the item "contributes to the attainment of some end or purpose of some user or system." Any artifact has an obvious function in this sense but, unless one believes in real goal-directedness in nature (as Wimsatt and I and only a few others seem to), organisms and their traits would not. This teleological sense of "function" is picked up again seriously in the goal-contribution account given by Boorse.

Lastly, in order to see why Wimsatt eventually groups all three senses, it is important to notice that both the evaluative and teleological senses can be seen as subclasses of the perspectival sense. If we say that something has a function in the teleological sense, we are taking the perspective of the user or system to whose end or purpose the function contributes. Having a purpose requires having a particular subjective perspective. If we say that something has a function in the evaluative sense, we are taking the perspective of one who defines a particular (subjective) standard of performance. In the end, a theory of function has to explain why functions are (i) perspectival, (ii) evaluative, and (iii) teleological; Wimsatt may be right, though, that all three can be explained by one model.

[146] Here, the word "proper" should be taken to mean "correct" rather than "property".

D. Improper Functions: The Illusion of Proper Functions

We say the soldier's Bible, on this occasion, performed the function of bullet-stopping . . . We do not say that stopping bullets was the function, or even a function, of the Bible, or that the Bible had this function.

—Christopher Boorse (2002)[147]

Unfortunately, philosophers' attempts to provide rigorous analyses of teleological concepts frequently make little progress against ingrained ways of thinking about these concepts.

—Marc Bekoff and Colin Allen (1995:253)

If we follow the past century of literature on the topic of function, we find ourselves interested principally in what I've just termed "property-functions". We find ourselves attributing a property to an item and then wondering what it is about the item itself (its history, its structure, its role) that allows it to have or to hold that property. In fact, as mentioned earlier, the term of art introduced by both Neander (1983) and Millikan (1984) and adopted by many other philosophers since is "proper function". The word "proper" in this context is used not in the sense in which it means "correct" but in the sense in which it relates to ownership—a proper function is a *property* of an item. (Although, as we saw with fringe function-attributions such as the ozone layer, not quite all property-functions are considered to be proper functions.) There is a deep assumption underlying this view that has gone largely unquestioned in this literature—the assumption that an item can in fact objectively *have* a function. I think this assumption is false. Things do not have functions.

[147] The ideas in this section expand upon and were inspired by similar ideas introduced by Christopher Boorse (2002:71).

Saying that nothing has a function is not meant to deny the existence of function*ing*, though. On my view, functions are just things that are served rather than had. They are properties not of an item, but of a relationship that an item may enter into, and their lifespans last only as long as those relationships do. Things *can* function, and things *do* function all the time, they just don't *have* unwavering intrinsic functions.

To get a first intuition for this idea, we can look at an analogy between function and the physical property of weight. Weight's close cousin, mass, can be called an intrinsic property[148] of an item—invariant throughout the universe (as far as we know)—while weight itself varies in some way with respect to the item's context (we can call it an extrinsic property)[149]. For example, a bowling ball may have six kilograms of mass and, here on earth, it correspondingly weighs nearly 60 newtons[150], but on the surface of Mars it weighs less than half that much while its mass remains unchanged. Furthermore, if we put that bowling ball aboard a spaceship bound from here to Mars, for most of the journey it will be relatively weightless despite not having undergone any change in mass. The difference here is that while mass is a property solely of a physical object, weight is a property derived from a *relationship* between an object with mass and a particular environment—in this case, an environment in which another massive object exerts a gravitational force upon the object.

[148] Though we must be careful not to confuse this intrinsic-extrinsic distinction with the identical terms that are sometimes used by physicists synonymously with the terms intensive and extensive properties. On the intensive-extensive distinction, mass is considered to be an extensive property, which means that it varies with regard to the quantity of material present (unlike, say, density, an intensive property which does not vary when a material is halved or otherwise divided).

[149] Extrinsic properties are sometimes called *relational* properties, though there is controversy over use of that term as a general synonym for extrinsic properties (Humberstone 1996).

[150] Six kilograms times approximately 9.8 meters per second squared of acceleration due to the gravitational pull of the earth.

While weight itself is not illusory[151], here on earth it does present a (typically benign) illusion that we could term "the illusion of weight *constancy*", a useful fiction by which we conceive of weights as fixed quantities because the context in relation to which they may vary—distance from massive objects such as the earth[152]—is so seldom varied that we imagine it to be universal[153]. These days, young children and other people with little or no training in physics or engineering are the most susceptible to this particular illusion, though any of us could fall for it at one time or another and, not too many centuries ago, perhaps most of us did.

Now what I am suggesting is that having a function is more like having a weight than it is like having a mass. It is a property that is not intrinsic to the items that appear to have it and it is subject to variation—functions may come and go just as weight can, though, as I'll explore momentarily, our very strong intuitions often seem to be that they cannot.

One reason I suspect that functions appear to be fixed is that our typical modes of thought tend not to vary the contexts in relation to which functional status might vary. We tend to think about traits mainly in relation to organisms and about artifacts mainly in relation to ourselves (as users), which are precisely the contexts in which such items are *able to* function. We seldom imagine a record player floating alone through the vastness of space after the human world has been destroyed[154], or a dismembered thumb that does not dream of reattachment. And even when we force ourselves to picture items divorced like this from their usual contexts, we can't help but continue thinking of the isolated item in terms of its usual contextualized behaviors. We can't shake the feeling that the thumb is for opposing the fingers in grasping, even when it is no longer able to

[151] It is an observable and measurable pattern in the world that, *ceteris paribus,* we can use to make reliable predictions (see chapter II).
[152] Weight also varies with other forces that act upon the same body—for instance, your buoyancy affects your weight when you are submerged in water.
[153] See also the illusion of color constancy discussed in Chapter II.
[154] The unmanned space probes Voyager I and II, which were launched in 1977, contain the needle of just such a record player, along with instructions for its use and a record to be played, in the unlikely event that, one day, the ship is encountered (and found to be relatively undamaged) by distant beings.

do so, partly because the thumb either does *that* or it does nothing at all to speak of. Aside from in its role in grasping, the thumb is an uninteresting object which we simply have no framework for thinking about, and the same goes for a record player taken outside of the context of a person using it to play records[155]. Since we tend to think of items only in terms of how and when they are able to perform their central functions, that functioning seems to be ever-present in the item, when in fact the ever-presence is only in our minds, in our experiences with, and our thoughts about these items, and this, I suggest, contributes in large part to our belief that their functions are enduring properties, a belief which (following the analogy from weight) we can term "the illusion of function constancy"[156].

The suggestions above help to partially explain why we might be liable to see constant functions *if* they didn't actually exist, but in order to fully expose the proposed illusion we still need reason to believe that functions may not actually be constant. We need to show that the functions of items can come and go. And while it should be obvious that, at some point in time, the material from which an item was eventually made did not have the item's function and that one day down the line, when the item disintegrates, that same material will also no longer have a function, those boundaries may be too extreme to convince a skeptic who might claim that functions may appear and disintegrate in tandem with an item's identity, but otherwise inhere in the item[157]. What we need

[155] Duncker (1945) named this bias "functional fixedness" after he showed its pervasiveness with his well-known candle-box problem in which participants in a problem-solving task found it difficult to think of a matchbox as anything but a box for holding matches in. Solving the candle-box problem required them to imagine the box being used instead as a small shelf (see also Frank and Ramscar, 2003). Birch and Rabinowitz (1951) revealed a similar functional fixedness bias using an adaptation of Maier's (1930, 1931) two-cord problem. In Birch and Rabinowitz's experiment, items previously used by the participants for a specific function (electrical switches and relays) seemed to be cognitively inaccessible for use in a novel way (*e.g.* as the bob of a pendulum) when later solving the problem of bringing two hanging cords together.

[156] A very similar illusion is the sense we get that a particular road or hotel or theme park or beach is always very busy, when, in reality, this sense is only because we are only ever on these roads or at those beaches at the times when everyone else also is, and so whenever we think of them, we think of them as we've experienced them: full of people (thanks to Seth Frey for pointing this out to me).

[157] Though, such a skeptic will surely be at pains to define what is meant by "an item's identity".

to show this kind of skeptic is that functions can come and go in items that remain largely or wholly unchanged.[158]

However, breaking the illusion of function constancy in our minds, while leaving functional items largely intact, turns out to be a tricky matter for at least two reasons. For one thing, our susceptibility to the illusion, along with the reinforcement it may receive from a linguistic community that has culturally codified the illusion, causes us to intuitively project functions (as enduring properties) onto items that behave functionally, or that are even just imagined to[159], and so we cannot trust our own instincts and intuitions about the matter. We cannot easily ask ourselves "Well, in a case such as X, does the item still have a function?" For another thing, there appears to be a chain of fallacies that collude to reinforce the illusion of function constancy. I will call the ones that I've identified by the following names: the design fallacy, the malfunction fallacy, and "for"-conflation. Before arguing that functions can come and go, first I will briefly describe these fallacies; detailed discussion of each will come later.

The design fallacy occurs when we believe that an item acquires its function through the process of having been designed. So when we imagine a dismembered thumb or a phonograph floating alone in deep space and we find ourselves convinced that these items still have functions, part of what convinces us of that is the idea that these things have been designed, in one way or another, to perform certain behaviors. We convince ourselves that such performances are what the items are *for,* regardless of whether the items ever actually perform them. Once impregnated with that function, the thinking goes, the item retains it permanently (at least, perhaps, until the item

[158] While use of the term "unchanged", here, ultimately needs to be understood in terms of my discussion of identity in chapter II, the standard interpretation of the term is a near enough approximation in most cases.

[159] Imagine for instance, an archaeologist or anthropologist discovering an item that has been carefully formed in some way but which had never been intended for something—perhaps it was the negative of a machined item, the scrap left over from creating something else. An object like this may have sharp edges, long lever-like members, evenly spaced holes, or various other marks of intention and function, yet it was never intended as any particular kind of tool. Upon discovering such an item, the explorer might venture some guesses as to what purpose the item served and, if a particularly fitting explanation seemed to account for all the features of the item, then it would be labeled as having been made for just that and no one would be moved to dispute the function attribution, despite it being wrong.

disintegrates). I must admit that, on the surface, this sounds more reasonable than fallacious, but I will examine (and try to dismantle) the intuition in a later section of this chapter.

The malfunction fallacy occurs when we believe that an item that is unable to function (either because it was designed or made improperly or because it was damaged in the meantime) still "has" the function that it is unable to perform[160]. For instance, a corkscrew that is badly bent to the point that it can never be screwed into a cork again or a congenitally deformed liver that cannot perform its metabolic functions are both considered to "have" the functions which we consider to be their malfunctions because after all, despite their failures to perform, that is what the items are "for", in some sense of the word (see, *e.g.*, Griffiths 1993; Millikan 1989a; *cf.* also the topic of "for"-conflation, below). The intuition can be compelling to some of us, but so can its opposite—that is, the intuition that an item that fails to function simply either never had or no longer has that function. A broken corkscrew can easily be thought of as nonfunctional or no-longer functional, rather than malfunctioning. I am, at times, convinced by each of these opposing intuitions. However, if we follow out the implications of the claim that an item can never "have" a function in the first place, then there simply is no question of whether that item might keep or lose that function when it becomes deformed (because there never was a function in the first place that might be kept or lost). *Both* intuitions would in fact be faulty. I'll argue against the malfunction fallacy in the next chapter when reviewing the SE analysis of function, whose proponents rest a series of their arguments upon the mistaken intuition.

"For"-conflation is closely tied to both the previous fallacies. The idea of a function (or a purpose) being what an item is *for* seems intuitive, but asking ourselves merely what an item is for

[160] This is closely related to the issue raised earlier about items (such as the record player in deep space or the dismembered thumb) that are unable to function when removed from their context of usage. There is a blurry boundary between them, of course. A malfunctioning thumb is, for instance, one that has been crushed badly enough that it can no longer function whatsoever in attempts at grasping, but it has not been removed. A decontextualized thumb is one that is no longer on a body even if, for instance, the entire forearm has been removed. But it seems there can be no fine line separating these kinds of categories.

causes us to carelessly blend or simply mix up a broad array of concepts, each of which is sometimes abridged in our mental shorthand to just its attendant preposition: the word "for". Even though the dismembered thumb is no longer *used for*, nor *good for* anything in particular (especially grasping), and though it is certainly not *there for* anything (imagine, perhaps, it is lying on the ground), and although having been constructed by natural design means that it never was *meant for* nor *intended for* anything (in the sense of foresighted human intention), we still might imagine it to have been *designed for* or *made for* grasping (in the sense only of the natural design of evolution). And so, while contemplating this particular thumb's function, when we ask ourselves "What is the thumb for?", an automatic assumption about just what kind of "for" we are talking about allows us to have an answer at the ready—an answer that, in the case of the thumb, happens to corroborate the design fallacy and make it easy to fall for the illusion of function constancy. We thereby believe that a dismembered thumb that cannot function in any way still "has" the function of aiding in grasping, and this, furthermore, buttresses the malfunction fallacy. If, however, we look at another case, say that of the phonograph, we can say that the item was *designed for* or *intended for* playing records, but only in the sense of human intention—human design, not natural design (cf. Lewens' Artifact Model, p. 224, below). And in further cases that I'll explore later, neither of the design-based senses of "for" can play a role in our belief that a particular item is for a particular role, but in those cases the notions of *used for* or *good for* or *there for* seem to fit instead, and so we often find ourselves justifying our expectations of those items having functions in those terms. The problem is this: if we must cite widely varying concepts across the various cases, then the apparent unity of "being for" something does not validly justify judgments of whether or not an item has a function . . . unless some broader, more generic, part of "being for" somehow serves to unify these other relations. I believe that is the case, but no plainly applicable taxonomy of "for" has yet been given to solve the problem. I'll review some ideas about the differing types of "for" in the next section.

At this point, keeping in mind the fallacies I've just alleged and the other concerns listed above, I can now attempt to show that functions are not constant properties held by items. The alternative that I will advocate is, as I said earlier, that functions are *served* by items when those items are in a particular context of usage, and so they are properties, instead, of that *relationship* of usage. When that relationship is commonly held in the world (for instance, hearts or eyes or thumbs that are parts of a body and that disintegrate relatively quickly upon removal from the body), or when that relationship is commonly held in our minds (for instance, hammers or computers or phonographs that are either thought of in terms of their being used by us or else not thought of at all), then we come to ignore the context of the relationship and carelessly think of these items, themselves, as independently "having" the functions that they regularly serve.

Let's look now at the boundaries of those usage contexts to see if we can find places where functions can seem to come and go from items that remain unchanged, while their contexts vary. As I mentioned, it is easier to look at items that seem to gain functions, since the cognitive inertia of the illusion of function constancy tempts us to continue attributing functions to items that might otherwise seem to lose functions. Both categories exist, but let's begin with the easier one.

Imagine you are just moving into a new home. During your first night there, while in the process of unpacking, you set a small empty packing crate on the floor near the fireplace and leave a pillow inside. Now imagine your dachshund chooses to sleep in this box (this scenario is drawn from Wright's example given above as FS13). In the event that something like this occurred, we would say that, for that night anyway, the box will *serve* the function of being a dog's sleeping quarters, but we would not yet say that it *has* the function of being a dog's sleeping quarters. It is a packing crate that the dog happened to sleep in.

If, however, you are too lazy either to move the box or to make a more proper bed for your dog, as the days turn into months and the months wear on into years, eventually we are willing to

say that the box now *is* the dog's bed and that it now *has* the function of providing a warm, cushioned, comfortable and familiar place for the dog to sleep. Nothing may have changed about the physical structure of the box, but time and repetition were able to transform it from being an item that we would say was *serving* a function—from verb-functioning as a bed—to being an item that we would say has a proper function, namely that of being a dog's bed.

One might interpret this to mean that the passing time somehow has the power to grant a function—to imbue an item with a function that was not originally there. We can keep that possibility in mind but I think it is at best imprecise, not the least because no meaningful physical changes have occurred to the box during that time. At worst it is an unwarranted hypothesis since we have no theoretical basis for suggesting that time alone has any direct causal capacity with respect to function. The interpretation that convinces me instead is simply that, seeing that the box served in this capacity—*served this function*—for such a long time, we fall under an illusion of function constancy, forgetting or ignoring that the box once was a shipping crate or that it could one day be one again. Aside from picking up a bit of dog odor and some slight wear and tear, the box is fundamentally unchanged in all meaningful physical respects, but what has changed is the purpose it has come to serve, the frequency with which it has come to serve it, and the *limits of our imagination* about what uses the box will serve during its existence. Although we could imagine something more if compelled to, without any motivation to creatively repurpose the box, and after observing it serving as a dog's bed for so long, we simply think of it only in terms of its serving as a dog's bed.

The key here, I propose, is that our attributions of functioning can vary with our perception of an item's *usage*. The more regularly an item is used for some purpose, the more likely we are to claim that the item "has the function" of serving that purpose. Notice that anything can be repurposed in this way—a rock, functionless at first, could (again, without physical change) make its way into a primitive toolkit as a hammer, or onto a desk as a paperweight (FS7), or into a foyer as a

doorstop. Despite having at first only functioned *as* one of these items, eventually—in a situation of increased regularity—we would call this particular rock *the* hammer or *the* paperweight or *the* doorstop and consider it to have the function that makes it what it is. Nonetheless, what changed is nothing about the rock, but only something about the regularity of the relationship in which it is used and, principally, our conception of that relationship as being *the* thing that that rock does or is for.

Cummins (1976) gives a similar example of a natural bowl-shaped depression in a large rock coming to be used one day to hold holy water in religious rituals. Eventually, if these rituals are held regularly enough (and perhaps if the rock is revered enough never to be used in other manners) we are willing to say that the depression has the function of being a holy water vessel. But when did it come to have that function, and what process or structure caused it to have that function? It seems to me that we assign it as having that function only because we observe the regularity of it serving that function.

The foregoing examples all involve items that seem to come to have functions without changing the item. Most of them are found objects—artifacts primarily in the sense that they have come to be employed towards some end, though none of them were really designed or made; they just happened to already have the required features that allow them to serve particular purposes. The reason biological traits are conspicuously absent from these examples is simply that, once development is complete, traits of organisms appear to already have their functions[161], and so it is less straightforward to look for a (biological) item that seemed not to have a function before coming to have one over time.

[161] That is, if the trait has a function at all. Spandrels—the space fillers that bridge the gaps between functional traits (*cf* Gould and Lewontin, 1979)—have no functions. But if they came to be used in some way (over evolutionary time, not an individual lifespan), then they would be termed functional traits (or *exaptations*; *cf* Gould and Vrba 1982) in their own right and thus would not count as functionless items at the scale of an individual life.

We can look now at the category of functions that seem to disappear in items without changing the item. The best examples, I think, are situations of abandonment. In the case of artifacts, we can imagine an archaeologist discovering a cache of fully preserved tools of an ancient civilization that were made for performing a job that we don't need to perform in the modern era, for whatever reason. Such tools might be labeled as having *had* the function, but no longer having the function. It is the context of those ancient people and the job for which they once used the tools that allows the attribution of a function, but in the modern absence of that context, the tools may be quite identical in form to their original physical form and yet no longer have the function of doing whatever it is they once did. In the case of organisms, we can similarly talk about items or traits such as molts. When a crustacean or an insect or reptile gives up its outer skin or carapace, or when a bird sheds a season's feathers for replacement, we are more inclined to think of these items as previously *having had* the functions they once served, but of no longer continuing to have those functions. They are no longer skins or carapaces; they are molts. Just as with the items discussed above that come to have functions, we consider these items first to have and then to lose their functions relative to their context of usage—when they can regularly serve a function, we consider them to have that function, and when they no longer serve in that capacity, we consider them to no longer have the function[162].

There are a couple of lessons I think we should learn from the recognition that functions can appear and disappear in items. The first point I find important is the surprising claim I began with: that functions are not real properties of items. *Proper* functions do not exist. Things don't have functions (TDHF)[163]. We should think of functional items not in terms of their *having* a

[162] There is an interesting contrast between our intuition that a molted skin no longer has a function, and our intuition that a dismembered thumb might still have a function, when both can be classified equally as body parts removed from the body.

[163] From here forth, in this dissertation, I will use the abbreviation "TDHF" to continually emphasize the claim that *Things Don't Have Functions.* The claim that things do have functions is so deeply embedded in the belief system of our

function, but rather in terms of their *serving* a function, though often doing so very consistently and reliably. This has broad implications for the philosophical analysis of function, since the modern debate has centered on the question of what the conditions are for an item to have a function. I take it that there simply is no answer to that question, and so the hunt is futile. Instead, we should try to understand what the conditions are for an item to serve a function and, perhaps further, to serve a function regularly, since it is such regular service that usually provides us with the illusion of function constancy that causes us to attribute or project proper functions. The second point I want to take away from the foregoing analysis is that since usage plays a role in serving-a-function and the frequency of usage often plays a role in having-a-function, we should keep in mind that perhaps some part of the concept of *usage* should play a theoretical role in the underpinnings of functioning. This will be the case in the theory I'll later advocate.

---

culture and in most of the other literature that I will be citing and arguing about, that I will need to constantly and repeatedly remind us not to be lulled by this natural assumption. When examining a theorist's example that, say, "a wing has the function of flying", we will need to take their theoretical point (whatever it may be) seriously while at the same time remembering that we are also dealing with a false claim. At such times, I will attempt to point this out by stating, "TDHF!" but I may also lapse into repeating such phrases as "the function of X is Y" or "X has the function of Y"; I ask the reader to take such remarks not literally but elliptically for something more along the lines of "we perceive X to commonly function in capacity Y" . . .

E. "For"-Conflation: Designed For, Used For, Good For, Meant For . . .

[I] am prepared, for the purposes of this paper, to let 'What is it for?' be a way of asking the same question as 'What is its function?'

—Robert Cummins (2002)

When developing intuitions about the subject of either *function* or *purpose*, an analyst is often inclined to ask themselves what an item is *for*, in the same way we did with the heart and the sailboat's standing rigging in the introduction to this chapter, and the same way we do whenever we contemplate the function of any item such as those in the list of function statements on pp. 205–207. In fact, many theorists that we will encounter in the next chapter have framed their inquiries into function in terms of the question "what is it for?" (*e.g.* Cummins 2002; Melander 1997:45; Neander 1991b:454).

Asking what an item is for may be unavoidable when thinking about functions but, as I pointed out while describing the illusion of function constancy, it runs us into trouble because the word "for" is often a linguistic shorthand that can be interpreted to mean a broad number of things—designed for (by blind natural selection), designed for (by foresighted humans), meant for, intended for, used for, made for, there for, done for, or good for, just to name the most common and general, purpose-related connotations[164]. These many interpretations of the word "for" reflect quite diverse concepts that are often not made explicit. When we are incautious in our thinking, we

[164] There are plenty of more specific kinds of purpose-related fors—too many to analyze them all—but most are subclasses of those noted. What is that pill *taken for*? (For curing bacterial infections—this is a subcategory of "used for".) What are spices *powdered for*? (For releasing flavor molecules—this is a version of "done for".) Who are playgrounds *built for*? (For the youth—this is some mixture of "made for" and "done for".) Similarly, the "for" in the sentence "She has an eye for fashion" can be replaced with "that is sensitive to", which means, roughly, that the eye appears to have been *made for* that sensitivity or at least that the eye is *good for* that (*made for*, in this case is used in the sense that it has been *trained for* it—a kind of *made for* that actually depends on molding the plasticity of the brain that stands behind the eye).

may find ourselves tempted to blend together, mix up, or opportunistically choose from these concepts, thereby possibly confusing any analysis that may depend upon precision in the interpretation of the word "for". Just such precision is required, however, in an analysis of teleological terms such as function, purpose, and goal-directedness since, as noted, we draw some of our central intuitions about these topics from thinking in terms of the question "what is it for?" (See also Dennett 2014:49-50; Nissen, 1993:19-20).

Later in this section, I'll describe what I think may be the most basic type of "for" in teleology—the sense of the word that I think is universal to all functional claims and that I suggest truly justifies the intuition that functional and purposive items are *for* something or other.

Lewens' Artifact Model

Perhaps the most important "for"-conflation to take note of is that which occurs between the two kinds of being *designed for* that lie at the core of what Tim Lewens (2004) calls "the artifact model of evolution". Lewens uses this term to refer to the way the analogy between human artifact design and the natural design of organisms makes it easy to treat the products of evolution as if they were artifacts. These two processes of design, as well as their products, share some important similarities (and thus a name) but also some important dissimilarities . . . centrally, the presence in human design, and lack in natural design, of intention or foresight (Allen and Bekoff 1995b; Lewens 2004: 115-116; Reiss 2009).

While using the artifact model in reasoning about biological items can at times be useful, even indispensable, it can also lead one to mistakenly treat those items as if they and their functional relationships had been produced by a conscious designer or as if those items are optimally designed for their role, when they often are not. Moreover, it encourages us to fall for the design fallacy,

coming to think that *design* might form the nexus between artifacts and organisms that contributes to their both being functional (see *e.g.*, Griffiths 1993; Kitcher 1993). I'll attempt to dispel that notion in the next section; for now, let's look at the two processes of design to see what they share and how they differ, and to get a handle on what theorists take design to be.

Natural selection is considered to be the main contributor to the gradual refinement of biological traits that is taken to constitute natural design. To review, natural selection is a process through which relatively more ineffective versions of a trait are filtered out (via relatively less successful reproduction of their bearers) while relatively more effective versions will survive more often, such that the offspring of the latter are more likely to comprise the next generation. Assuming a source of variation in the population (and assuming the selection pressures remain relatively unchanged) iteration of this process over many generations is able to create progressively more effective versions of a trait, and this progress is what is usually meant when people talk of "natural design". Because of the iterative, cumulative, trial-and-error nature of the process, natural design is often deemed a "generate-and-test" procedure, highlighting the two main factors in the process: construction and evaluation (Dennett 1995; see also Allen and Bekoff 1995b).

Now we all understand the process of human design: with widely varying degrees of engineering work, an idea is converted to a prototype through some kind of construction process. The prototype is then assessed to see if it works, after which modifications are iteratively made to the prototype until the designer is satisfied with the behavior of the item. For obvious reasons, the iterative process in the human design of artifacts has been compared to the generate-and-test paradigm of natural design. Theorists note that in human design the generating and the testing may occur in a few ways. First, it may sometimes occur in the world over cultural timespans, as only the more useful artifacts, or the most useful versions of them, are selected to be copied and gradually modified by a culture. Second, it may sometimes occur in the world over briefer timespans, as a

designer generates and then tests model after model, over the course of days, weeks, or months. And, third, it may sometimes occur in the mind of a designer, as modifications to an item are imagined rather than built, and their performance is *mentally* evaluated before being either forgotten or ultimately built for real-world testing (Dennett 1995, Griffiths 1993).

Adaptationism

Despite the fact that one process is driven by human intentions and the other is driven by chance, it is clear that human and natural design can be seen in some regard as quite the same procedure of generate-and-test. Indeed, viewing them as the same in this way has led to one of the most fruitful courses of reasoning in biological study—the reverse-engineering or adaptationist stance that allows us to approach biological items as designed items in order to figure out how they work or what they are for (see, *e.g.*, Cosmides and Tooby 1992; Dennett 1995; Hurley, Dennett and Adams 2011, 2013; Krebs and Davies 1997; Lewens 2004; Pinker 1997; Williams 1966).

Dennett (1983, 1988, 1990, 1995) is perhaps the most vocal proponent of using the artifact model in the reverse-engineering of biological traits.

> Instead of trying to figure out what God intended, we try to figure out what reason, if any, "Mother Nature"—the process of evolution by natural selection itself—"discerned" or "discriminated" for doing things one way rather than another. (Dennett 1995: 213)

This adaptationism has been mocked by others in the field who point out that the adaptationist's working assumption—that Mother Nature has designed things for reasons—seems to require a false

belief . . . namely that there is an agent we can call Mother Nature and that she designs solutions in order to solve certain problems (*e.g.*, Gould and Lewontin 1979; Lewontin 1993)[165]. But proponents of adaptationist thinking generally don't agree with this caricature. They argue that the adaptationist thinker doesn't need to *really* believe that Mother Nature is a foresighted agent nor that "she" creates optimal solutions[166] nor that every trait they are interested in was designed; instead they may simply proceed in their inquiry *as if* the products of natural selection had been designed, regardless of how they came to be. That is to say, the analogy underlying the artifact model can be used heuristically, based upon the resemblance between the *products* of evolution and those of human artifice without taking seriously the notion that the *processes* that produce biological items fully resemble those that produce artifacts (Lewens 2004; Resnik 1997).

As-If Reverse-Engineering

The question that some adaptationists may be asking, and the question that anti-adaptationists rightly rebel against for its design implications, is "what is it designed for?" Whether or not one differentiates between the two versions of design, this shouldn't be the first question to ask when approaching a reverse-engineering task because it asks us not only to hypothesize how the item functions but also to hypothesize how the item historically came to be functional, under the assumption that the latter explains the former. That is, the problematic version of adaptationist reverse-engineering assumes that an item is functional *because* it was designed to be, and so, to divine an item's function, it requires reasoning about the item's design history (though remember: TDHF).

[165] The critique is richer than this, but the minutia of the debate over adaptationism goes beyond the scope of the current project.

[166] Lewens (2004: 47) points out that the failure of optimality assumptions applies equally to reasoning about artifact design. Nothing about artifacts requires that they be optimally designed and, usually, they are not. But we can still reverse-engineer artifacts by reasoning about what their various parts are for.

Examples where this assumption is false have now been widely explored. The classic illustration is how the feather, which is functional in modern birds for flight, originally evolved on flightless dinosaurs, presumably for thermoregulation or possibly courtship or aggression displays (Gould and Vrba 1982; see also Ji and Ji 1996). What feathers and many of their detailed features were (naturally) designed for is not the only thing they function to do today, and, despite our not knowing about their design history until recently, we were still able to sort out their function in flight.

There is another question, though, that does most of the work of "what is it designed for?" without making any historical assumptions. If we relinquish our commitment to identifying function with *design* (Allen and Bekoff 1995b) and ask instead "what is it *good* for?" we have a presently answerable question about an item's functioning that can help us look for evidence about the ways an item is actually able to function (Amundson and Lauder 1994).

Of course, once one has answered "what is it good for?" not only does one already have an idea of how the item in question functions—that is, one has completed their reverse-engineering—but one can then *secondarily* conjecture, albeit fallibly, as to how the item may have been designed for the behavior which we've discovered it is good for and, presumably, if one were so inclined, one could then attempt to inquire into historical facts for evidence supporting those secondary, design-related hypotheses (Hurley, Dennett and Adams 2013). It's not that we can't reason about a functional item's history and its design, but that the inference runs in the opposite direction—from function to design rather than vice versa.

Other Kinds of For

Earlier, I examined the examples of a dismembered thumb and a phonograph far from any user, noting then that our intuitions about function in each case are based on differing senses of

having been designed for (if you recall, these examples were chosen because, in their hypothetical contexts, they are actually unable to function). If being naturally designed for something and being humanly designed for something were the only two interpretations of "for" available then we might be convinced to subsume them under the more general generate-and-test concept of design. However there are cases of items that do function, in which our sense of "what they are for" may seem as if it's neither of the design-based senses.

Take for instance the case of someone building a swim dock at their lake house. Imagine that, in an environmentally conscious effort to recycle materials, the handy homeowner constructs a wooden-framed deck into which they then pack hundreds of used plastic soda or water bottles for flotation. In a case such as this, the bottles were neither humanly designed for this purpose nor naturally designed for this purpose. But when we ask what the bottles are for in this dock, we will say that they are for flotation and what we mean by that is that their function in the dock is flotation. We have a function for which design, construed as generate-and-test, played a very limited role, if any at all.

What may instead justify the use of "for" to describe the bottles' functioning as flotation is that they were *intended for* the job, they are *good for* it, they are *there for* it, and they are being *used for* it. There is of course a sense in which design played a role in the bottles coming to be floats—the dock-builder intended them to serve in that manner as a part of the dock. But the bottles themselves were neither made of the material they are made of nor molded into their form in order to serve this function. They were designed not for holding air in and water out but just the opposite.

One way to try to salvage the design-based interpretation would be to say that the bottles were designed for that function by a single round of generate-and-test. The dock-building homeowner had the idea to use the bottles, tried them out, and *since* they successfully floated the

dock, rather than iterating the search for a better design, the homeowner just continued to use them. I think this is a fair characterization but what it points at is that the core part of being *designed for* some behavior or other is really not in iteration or accumulation but merely in the *test* part of generate-and-test . . . it is in the determination of whether the item in question is *good for* that behavior.

What Is It Good For?

In analyses of function, analysts have taken up various interpretations of "what is it for?" For instance, Kitcher (1993) seems to identify "what is it for?" with "what is it designed for?" while Wright (1973) takes it to mean "what is it there for?" and Bedau (1992) thinks the point of interest is "what is it good for?" Other theorists are unspecific (*e.g.* Cummins 2002).

As one can see from my explorations above, I'm inclined to agree with Bedau. I think the many interpretations of "for" are all based on a core consisting of *good for*, as we already found with the two flavors of *designed for*. But before mining for those additional intuitions, first we need to distinguish between two further senses of the notion of *good for*. What I will call the personal sense of "good for" is the one in which an item brings *benefit* to a beneficiary: The measles vaccine is good for children. A little sleep would be good for me. This is the sense which Bedau uses, and it can be put in contrast with the impersonal sense of "good for" in which an item is particularly *suitable* for a certain task: The pen is good for writing. The screwdriver is good for prying. The discarded soda bottles are good for flotation. The difference between the personal and the impersonal is between being good *for someone* (providing benefit) and being good *for something* (being well suited).

There is a relationship, though, between these two senses: Whenever something is suited to a particular task, the fact that we think of it as impersonally good for that job derives in part from its

causal capacities but also in part from the fact that that job is personally good for some agent who benefits from its performance. An enormous meteor is particularly *capable* of extinguishing life on a planet but we wouldn't say it is *suitable* for that. Despite being *able* to cause mass extinction, there is no sense in which we would say this is *good* (except in the bizarre hypothetical in which someone had massively genocidal intentions; but the exception just proves the point). On the other hand, the bottles are good for flotation, the screwdriver is good for prying, the pen is good for writing, and the phonograph is good for playing records, whenever there exists someone who has the intention of using these items for these actions . . . that is, whenever these actions are beneficial to or (personally) good for that person. Similarly, the wing is good for flying, the eye is good for seeing, and the thumb is good for grasping whenever they are able to be used for these things by the beneficiary organism that owns them. But the thumb we removed from a body earlier is no longer good for grasping, since, after all, it is no longer able to.

Ultimately, what I am suggesting is that hiding behind every version of "what is it for?" is a *beneficiary*—the subject *whom* it is for. Anything that is personally *good for* some agent has a direct beneficiary; anything that is impersonally *good for* some activity is so suitable because that activity is good for some agent; and anything that is for something in any other manner—(naturally) designed for, (humanly) designed for, used for, meant for, intended for, made for, there for, done for—is, in one way or another, good for whatever or whomever it is for. To confirm this impression, we can look at each of the cases.

Our analysis of the impersonal sense of *good for* raised again the issue explored in the previous section about usage of an item: when something is suitable for something it is suited to be used, and to be *used for* something is to bring benefit to a user by serving them in one way or another.

We saw earlier that being *designed for*, whether it is human design or natural design, is to go through a process, which, crucially, contains a step (testing) that determines whether an item is *good (to be used) for* whatever it is that we consider the item to be designed for.

To be *meant for* or *intended for* something is central to human (though not natural) design, but these terms extend to purposiveness that goes beyond the design of artifacts as well. Whether or not it succeeds, a particular action, such as throwing, might be *intended for* creating some effect or *done for* the sake of that effect; a more abstract artifact, such as a message, may be *meant for* impacting a particular audience. In all cases, though, these performances are *good for* something or other—they are intended or done in order to achieve (or at least to create a chance of achieving) a benefit their performer perceives.

At a different margin of the notion of design, a part of an item, such as the primer bulb on a lawnmower engine, can be said to be *there for* whatever role it plays in the whole item (Cummins 2002) but that role requires the part being impersonally *good for* playing that role, and it only makes sense in terms of the overall item's being *used for* whatever it does. Alternatively, a whole lawnmower in the gardener's shed might have been placed *there for* the gardener's ease of access. This sense of being *there for* is also purposive, but the meaning is closer to *meant for* or *intended for* than it is to designed for or used for. Still, the lawnmower's placement is rooted in its being *good for* the gardener.

At the end of the day, in any case of any thing (an arrangement or intention or action or item) being for something or other, we can find a subject whose benefit from that thing underlies the conception of the thing as "being for" something or other. These brief explorations may not be strong enough to win over the skeptic of conceptual analysis, but the challenge, in order to discount these intuitions will be to find an example in which some version of being good for, meant for,

designed for, or used for, does not have a subject, implicit or explicit, whom the item, event, or action is ultimately for, in one way or another.

It is my contention that if a theory of function is understood to be taken in terms of the question "what is it for?" (as Cummins put it so directly, in the epigraph to this section) then this pattern of being for some agent, is what the theory ultimately needs to explain. Furthermore, I suggest we will become sorely misled if we attempt to interpret "what is it for?" in terms only of being designed for (Kitcher 1993) or of being there for (Wright 1973) or, worse, in terms of differing senses of being for, when analyzing different cases. The notion that ties these all together is that things that are for something or other are good for someone or other.

The wrench in the works that will have to be worked out first is that the notion of benefit or "good" is, at this point, entirely undefined.

F. The Design Fallacy

If an artifact was explicitly designed to do something, that usually determines its function, irrespective of how well or badly it does the thing it was supposed to do.

—Larry Wright (1973)

Roman coins, even if they are in mint condition, are no longer money, that is, they no longer have the function of being legal tender.

—Peter McLaughlin (2001:48)

Objects identical to our knives and saws would have no function at all if produced by some random process on a planet devoid of life.

—Christopher Boorse (2002:68)

Shortly ago, I alleged that we are susceptible to a number of errors in the ways we conceptualize functions. I called one of those errors "the design fallacy" and described it as the tendency to believe that an item has its functions by virtue of the fact that it was designed. In other words, we are often led to believe that design may grant functions to items. Not only is this a common intuition (see, *e.g.*, Allen and Bekoff 1995: 614; Griffiths 1993: 418; Millikan 1984: 17; Williams 1966: 9; and Wright 1973: 146), but it has also been offered as a theory of function (Kitcher 1993). In the previous section, I began to argue against the role of design in function by

drawing attention to the differences between natural design and human design. I will continue with my attempt to nullify that intuition now, in order to help us avoid theories of function that I think are wrongly based in design and also in order to bolster my earlier argument against the existence of proper functions.

The faulty intuition seems largely rooted in two seemingly evident facts. First is the perception that every place where we find functioning, we see that the items were designed, and every place where we see designed items, they appear to have functions—and this seems to be the case for artifacts as well as organisms and behaviors. Second is the perception that artifacts appear to be designed for their functions and that creating functional items is in fact *what design is*. Why else would we design something if not in order to create a functional item? This second notion comes with a biological extension, using Lewens' artifact model, whereby we presume (carelessly) that organisms are also designed for their functions, despite our being aware of the fact that modern evolutionary theory clearly understands natural design to be an impersonal process in which items simply cannot be designed to perform a preconceived function (see also McLaughlin 2001; Ruse 1981).

I will argue against both perceptions below. My conclusion will be that design never grants functioning, and that our conception of the relationship between design and function should be inverted[167]. I'll suggest that while design often plays a role in *constructing* items that are functional, it is the possibility of a relationship of function that is actually prerequisite to the process of design rather than vice versa. It is the normative role that *successful functioning* plays (as a stopping-point in the process of designing) that underpins the relationship between functioning and design. Items are functional because of the fit between their individual forms and a particular user and context of use. That relationship can exist whether or not design occurs.

[167] See also Allen and Bekoff (1995a, 1995b) for arguments against the equivalence of design and function.

If a norm of functioning were prerequisite to designing (as I am proposing), then we would expect to see situations in which design may be attempted but fails to meet those norms, resulting in incomplete designs that fail to function. We would also expect to see situations in which functioning occurred naturally or simply without requiring design to achieve them—situations in which the functional norms are met (somehow) without the effortful intervention of a designer. Examples will show that both these categories exist—that the correlation between design and function is incomplete—and this should lead us to infer that rather than design having the capacity to grant functions absolutely, something more like the notion just described takes place: functioning is a prior norm that plays a role in guiding design toward the construction of items that seem to have functions. This would still explain the regularity with which function and design appear together, but it would also explain the exceptions in which they don't.

Under this view, design doesn't guarantee or grant anything, but it increases the likelihood of an item becoming functional. It helps to ensure that an item will have certain regularities in its form and behavior that will lead us to conceive of those regularities as the nature of the item and that will keep us from conceiving of the item in other ways (Birch and Rabinowitz 1951; Duncker 1945) thereby coming to consider the item (*e.g.*, via the illusion of function constancy) to have a function.

Counterexamples

Design and functioning do regularly appear together but, for one thing, correlation is not causation and, for another thing, as it turns out, there are systematic exceptions to the regularity. We can find examples both of things that are designed but that do not clearly have functions[168], and

[168] Of course, I am granting too much here, since things never have functions (TDHF).

of things that seem to have functions but were clearly not designed, making the conclusion that design grants function tenuous at best (see also McLaughlin 2001; Sorabji 1964).

We can begin by recalling the examples from earlier: The dog's sleeping box, Cummins' depression in a stone used as a holy-water vessel, bottles used as flotation, and rocks used in various capacities as hammers, paperweights, or doorstops. In each of these cases, design (understood as generate-and-test) plays no role in the item either coming to be functional or coming to be construed as having a function. Allen and Bekoff (1995b) also give the examples of driftwood, seashells, and a taxidermist's preserved grizzly bear, all used as decorative items. Items such as these serve aesthetic functions without having been designed for such aesthetics (they were simply chosen; in many cases, merely collected from the shore)[169].

Conversely, there are items that have been designed but do not function, though making this clear is much less straightforward. One example would be the oodles of food products that are manufactured every year but that go forever uneaten as they pass their expiration dates and eventually get sent to landfill. Similarly, there are masses of electronic products that remain unsold, on shelves, well beyond the time at which they become antiquated or made obsolete by newer models. None of these products ever serves a function. As I've mentioned, though, it may be difficult for us to overcome our prejudices in cases like these—the intuition that these items have a function despite not serving it is strong, since the illusion of function constancy and the design fallacy and malfunction fallacy are so deeply rooted in our minds and our culture. To try to appreciate design without function, it is probably best to look instead at items that have not been successfully designed, since well-designed items typically succeed at functioning, making it difficult to see them as non-functional.

[169] It is worth noting that all of the various kinds of examples just cited have, instead of a "designed for" relation, a "used for" relation of some sort. Being used in some respect appears to make an item functional. Being conceived of only in terms of that use appears to make an item have a function.

Millikan borrows an example from a poem called "The Engineer" (Milne 1927) in which the character Christopher Robin has designed a brake for a train made from "a string sort of thing" and claims, despite its failing to smoothly stop the train, that "it's a very good brake" (Millikan 1989a: 296)[170]. With this quote, Millikan wants to support both the design fallacy and the malfunction fallacy; she wants to say that Christopher Robin is right, it *is* a brake—the item has the function of braking despite being unable to do so, and this is *because* that is what it was designed for[171].

What Christopher Robin is really doing with his claim, however, can be classified as metonymy or metaphor. He is alleging that in his own opinion it is a very good *brake concept*—a very good *idea* for a brake. The entirely nonfunctional so-called brake that he is holding in his hand, and apparently referring to, is cognitively standing in for the concept of a particular abstract brake design, which is the true referent of his remark. The so-called brake itself is not a very good brake, nor is it even a brake at all; it is a piece of string that does nothing of the braking sort, and so, taken literally, Christopher Robin's claim is false. He and Millikan are of course both aware of this—the sentence can still be true as long as we realize it is not a claim about the actual string sort of thing but instead a claim about a concept. But still it remains to be seen whether or not any brake of the kind that Christopher Robin has in mind would be functional in the slightest, much less "very good". If none of them will ever be shown to be functional—if Christopher Robin's idea fails to capture the physical essence of braking in any way—then it would simply be improper (however convenient) to use a functional category ("a brake") to describe any of them. If we want to use the category "a brake" to describe things that are simply meant to be brakes, or things that we hope will

[170] This is how Millikan quotes it, though if we look back at Milne, the actual quote is: "It's a good sort of brake, but it hasn't worked yet" (1927). The discussion here is about Millikan's use of her version of the quote, so I will follow that, ignoring Milne's original text.

[171] The problem that spurs Millikan's musing on the string sort of thing is described in this quote which precedes her text on Christopher Robin: "For example, exactly what sorts of (current) properties must an item have in common with some functioning token or other of a can opener in order to count as a *can opener that doesn't work*? The question is absurd on its face" (1989, italics added). My answer to the question is simple: None. A can-opener that doesn't work (which is to say, an item that is not a can-opener) need not have any properties in common with a can-opener that does work.

perform braking, or things that represent a particular brake concept but are not actually functional as brakes, that is a fine use of language, but we must be aware of the metaphor we are using and vigilant with ourselves not to allow it to influence our theory of the functional category of *brakes*, let alone our more general theories of *function* and *design* themselves, which may be derived from or colored by our intuitions about such example categories.

My claim here opposes Millikan's intuition: Christopher Robin's string sort of thing does not have the function of braking (primarily because TDHF, but also) because it simply fails to function as a brake, and we haven't yet any reason to believe that it in fact "is a brake". To the extent that anything Christopher Robin did resembled designing (and Millikan would indeed have us imagine a world in which his process would be considered design), that process neither made the string sort of thing functional nor christened it into the functional category of "brake" nor gave it a function. It is just a string sort of thing that we would be mistaken about if we called it a brake.

Allen and Bekoff make the following claim, similar to Millikan's.

> Prior to the 1903 Wright Flyer, many contraptions were designed for heavier-than-air powered flight, yet none of them flew. Modern aviation did not have to get off the ground (pun by design!) for it to be the case that the function of those remarkable contraptions was to fly. It was their function to fly because that is what they were designed (albeit poorly) to do. (Allen and Bekoff 1995b: 614)

Although I agree that this could seem intuitive, the same considerations as with Christopher Robin's so-called brake hold. These were not flying machines; they were *so-called* flying machines, *hopeful* flying machines. In this case, the word "function" is standing in for the concept of "intended function" or "hopeful function". Certainly it doesn't mean *successful* function. It is better to say that

the *intended* functioning of those remarkable contraptions was to fly. Allen and Bekoff's intuition seems to imply that intended function—merely wanting something to fly—constitutes designing something to fly. On that view, as long as they were intended to function in flight, they will have such a function.

Surprisingly, the same prejudices don't appear so strongly when we look at organismic traits, and I suspect this is because when reasoning about organismic traits, we don't bring to the analysis any sense of intention. In the case of organisms, when a new item does not work, we don't take the item to be a designed item; we take it to be a failed byproduct of the design process, while its competitors that *do* work are the ones that, we would say, have been designed. We don't say that a wing without feathers has the function of flying though it simply can't do so; we say that it is the result of an adverse mutation that destroyed in it the function of flying, which its ancestors had.

Allen and Bekoff's intuition that design grants function seems to rely on a belief that intention is sufficient for human design to grant functions ("It was their function . . . because that is what they were designed . . . to do"). Another view might suggest that another feature of design (say, generate-and-test) grants functions (even if the item fails to function) [172]. In order to adjudicate between these possibilities, it is worth looking at what roles intention and generate-and-test may play both in the process of design and the process of items coming to be functional. Let's do that now.

Does Design Construed as Intention Grant Function?

If design is merely intention, then an item need not succeed at functioning to have been designed. We could make attributions such as those from above: hopeful-handbrakes that don't halt anything and hopeful-helicopters that can't hover can still have functions. But how far can we take

[172] I'll also address these issues again in the next chapter when I discuss the malfunction fallacy.

this idea? Can I say that a six-year-old's drawing of a fantastical airship was designed to fly and therefore the model made from it "has the function of powered flight"? Can I present my own design for a perpetual-motion machine and claim it has the function of producing infinite energy? If I say I've just designed this piece of origami now sitting upon my desk such that when I pull on two opposing tabs, it will open up and swallow the earth . . . does that mean such a piece of paper *has the function of swallowing the earth*? Intuitions about these cases might differ from those about our earlier examples (handbrakes and helicopters). One might respond, "Well, those aren't *serious* designs!" but it seems as if the line between serious and non-serious designs will be difficult to draw. As the six-year-old grows and becomes more and more sophisticated about aerodynamics and buoyancy and so on, their designs will eventually come to resemble those of 19th-century "airplane" designers who were very serious but nonetheless failed in all their attempts to design and build functional flying machines. It is unclear how we would distinguish between serious designs that grant functions and silly designs that don't.

In any event, the idea of a "serious design" wouldn't get us any further. Take, for instance, a pair of scissors that cannot yet cut paper until we adjoin the two blades with the screw that acts as their fulcrum. This can be called a serious design not least because we know it will eventually successfully function to cut paper. But when does the design-minded theorist claim that these scissors come to have their function? If function is granted by the intention component of design, then it seems the parts need not be assembled since they and their interactions have already been seriously designed well before assembly. If that is our principle though—if an item that is not yet able to function can have a function as long as it was intended for it—then we can imagine the metal ingots that are destined to become those blades and that screw already have their functions before they are forged and cut. This begins to sound absurd, and it sounds even more so when we back the

clock up further to when the metal is distributed in ore somewhere in a future iron mine or when it is being fashioned from lighter elements in a nearby stellar fusion reactor.

Intention alone seems insufficient to constitute the kind of design that is imagined to grant function (not to mention TDHF) and, furthermore, if intention did play a role in granting function in artifacts, then the design-minded theorist would have trouble extending the notion to explain function in organisms without resorting to the intentions of a divine creator.

The Design-and-Construction Gradation Problem

We can call the above absurdity about when the iron that becomes a pair of scissors comes to have its function by the name "the design-and-construction gradation problem", and it may be useful to state the problem in other ways. Anytime an item is created, if it is to have a function in black-and-white terms, then we face the question of just when it comes to have that function.

If one held a theory (or even an intuition) that design granted functions, then in justifying that belief, one would ideally like to know just what part of the design process performed this christening. It would be meaningful to ask just how and when a designed item gets its function—by what means does design grant function? It would also be meaningful to ask by what procedure, over the course of an item's existence, it might lose its function. Certainly there is no reverse process—"undesigning"—that may eliminate a function, though "unbuilding"—disintegration—may[173].

I just argued against the proposal that design might be designated as the intention behind an item. Perhaps what matters, then, is the process of turning an idea or intention into an actual physical item? But if we must build the item, then when in the process of construction does it get its

[173] Perhaps, however, a look at the types of events that do take functions away can help clarify what is central to functioning.

function? When its prototype is complete? When the item's materials have been apportioned and reserved for it? When all the parts have been prepared (as in the disassembled scissors)? Or when the item itself is complete? And, besides, by what standard are we to judge the completion of any particular stage?

The central problem here is that functions had (or granted) are taken to be binary properties while the items that supposedly eventually have these functions go through gradual, fractional, piecemeal processes both in their design and their construction (not to mention that design and construction are often simultaneous and inseparable processes). It seems to me that the only way to make sense of this disparity would be to mark functions as existing upon some kind of completion standard—an endpoint of some sort—since finding a threshold in the middle of the process seems hopeless.

Imagine someone were designing a bladder for carrying a gallon of water to a location that requires over an hour of walking in order to get there. Imagine their first model is made of sewn canvas (call it Mark I) and that it ends up holding the water for three seconds before deflating. Would we say that Mark I has the function of holding water for over an hour? I think we wouldn't. Now imagine that, upon the discovery that it didn't work, the designer gets the idea to coat the material with a rubber of some sort before sewing it. Mark II holds the water for about a minute—it leaks out through the seams only. This is better than Mark I, but not nearly what the final product needs to do—so does Mark II have the function of holding water for over an hour? Still no. Finally, our water-toting designer gets the idea to put their rubber compound into the seams before sewing up the layers of fabric, and this happens to seal up the device so that Mark III indeed holds a gallon of water for many hours. At this point I think we would fairly say that the item has the function of holding water as hoped. But as we look back on the process, it is clear that what was required in order to finally christen the item with its function was a successful comparison to a

metric—a normative standard of functioning by which the product could be measured and the iterative process of design could be halted.

Does Design Construed as Generate-and-Test Grant Function?

One answer to the scissors question is to claim that function is granted at the moment when that screw is fully tightened into place fixing the relation between the blades—when construction of the item is completed and the design is finally and fully implemented. If this were the case, though, how would we determine that the design is fully implemented? What marks that threshold?

One way to know would be to see if the item reflects the idea or the intention by physical specification—that is, if it looks like the blueprint in all measurable ways. But if that were the case, it would mean little more than a return to the (unsuccessful) intention-centered theory of design or else a joint-criterion theory in which both intention and construction to the (physical) specifications of that intention together grant function. Fully assembled versions of the six-year-old's airship or of my perpetual-motion machine or of the earth-swallowing origami would have functions as long as they were physically built to spec. Unless we are prepared to grant functions to all fantastical ideas that don't work, we need to abandon the notion that this kind of design could grant functions.

Another way to say that the design is fully implemented would be to see if it *does the job* intended in the idea . . . that is, to ask whether or not the item measures up to the *functional* specification, rather than the physical specification, of the blueprint. This sounds reasonable, and it also prods us to consider design as a more dynamically normative procedure than in some of the foregoing analyses—something along the lines of a generate-and-test procedure. This kind of design can be seen as the refinement of an idea or a prototype until something that began as a rough inkling in the mind becomes an actual artifact in the world. Checking a model against the functional

specification acts as feedback that drives the design process. A scissors design would not be complete until the prototype was able to cut paper; a winged device would not be considered a flying machine until it was able to fly; a pouch would not be considered a water bladder until it was able to carry water. I submit that the failed machines that Allen and Bekoff label as "having the function" of flight would not have been so labeled by the engineers that were building and testing them, because they would not have considered their work (of building a flying machine) to be done until testing succeeded[174]. An item that has been designed but that does not function will be either discarded or designed further until it does function, at which point it will then fit its intended functional category.

It seems reasonable to say that design construed as generate-and-test can grant function, when *both* parts of that formula have been performed. But here's the rub: if successful functioning is the norm by which design is measured, if the final state of *being able to function* marks the end of the design process, then design itself must be defined in terms of functioning, not vice versa. Generate-and-test relies upon an external metric for measuring functionality—a metric that exists prior to, and thus needs to be defined independently from, design. In human design, that metric may be the idea or intention to which we compare an item before proclaiming it complete. In natural design, that metric is something along the lines of successful survival and reproduction. Because of the dependency of design upon functioning, it would be circular to consider function to be granted by design.

[174] They might have considered their work of building "Mark I" done when they completed this particular model. They might even have boldly claimed, "This is my flying machine" before a test flight, but the uncertainty in their hearts would have kept them aware of the optimism and imprecision of that statement.

Designed Not to Function

As an interesting aside, one may design items specifically such that they are not able to function. A number of artists have done just this, designing items that resemble familiar functional items but with single minimal changes that render them nonfunctional. Figure 4.1 show some examples of works made by the contemporary Greek artist, Katerina Kamprani. A watering can that cannot water, a fork that cannot skewer, intolerably slow shakers, pot handles that cannot lift, a broom that cannot sweep, and so on . . .

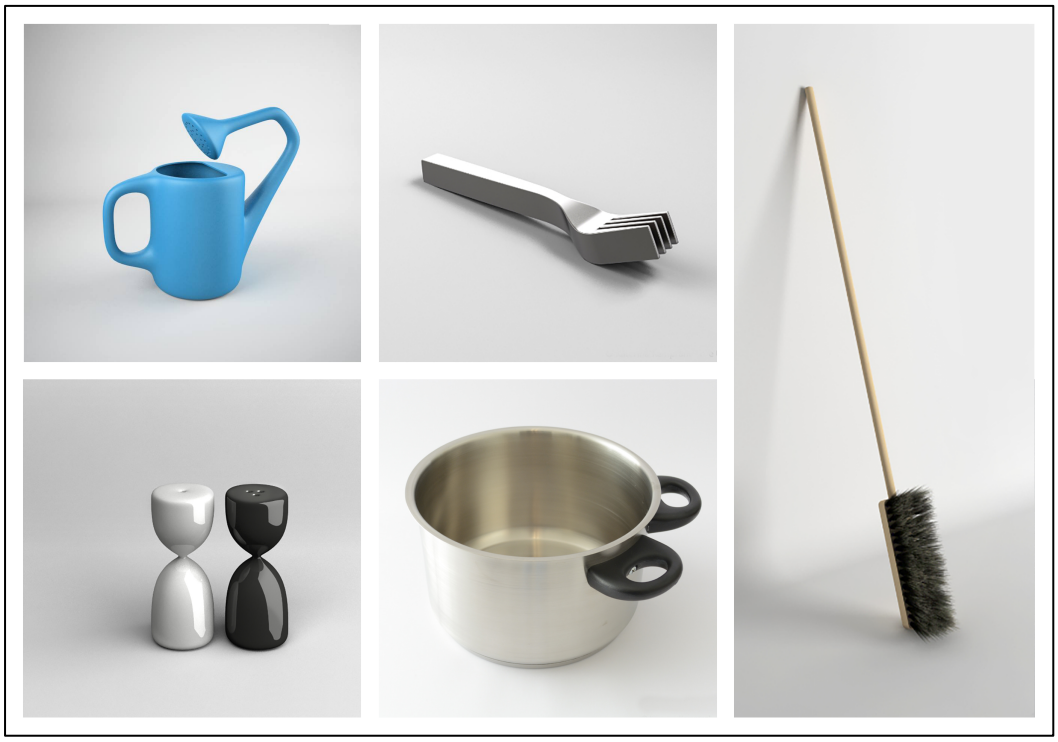


Figure 4.1: Objects designed not to function how we might want them to (reprinted with permission from the artist Katerina Kamprani).

Now these items have been designed to function as pieces of art—to provoke thought—not to function in the ways viewers might imagine they ought to function, and so they are not

counterexamples to the idea that design might grant function. But they are interesting, nonetheless, as they bring to light the fact that, as actual designers of these items, we would not consider them to yet have their functions until they were completed—until (at least) one more iteration in the design process developed them to a state in which they were able to function. Design would not be complete, in these cases, until our models matched up with functional norms.

There! It Works!

When we are designing a new item, we fiddle with materials and put them together in various ways, guided by varying degrees of theory and engineering knowledge. Sometimes our first attempt works, sometimes it takes iteration after iteration to get to what we want, and sometimes no attempt ever works. But either way, we know when we've reached the crucial stage in the design process when suddenly the item does what we wanted it to. The inventor is rewarded when their invention succeeds. And when the contraption (which the inventor might have been depressed about for weeks because of its resemblance to a piece of junk) finally, with one last tweak, succeeds, the declaration usually uttered is something along the lines of "There! It works!" What that utterance means is that the item suddenly has the form that gives it a functional capacity in a particular context. When it is tested, when it is used successfully in a particular way, when it shows itself to have certain capacities that allow it to be used in that way, then we recognize that, at last, it has become functional. Surely the designing and tweaking process contributed to giving it that form. But it is the form itself—the causal capacities in the item itself (in the context of a particular user and usage scenario)—that are recognized to be functional, at long last.

Natural Design

Much of the foregoing analysis focused on artifacts rather than on the traits of organisms. That bias occurred largely because I have been analyzing examples borrowed from a literature that, whenever discussing these intuitions about design, has focused solely on intuitions about artifacts (see, *e.g.*, Allen and Bekoff 1995: 614; Griffiths 1993: 418; Millikan 1984: 17; and Wright 1973: 146)[175]. And the bias within that literature, itself, occurs partly because *intention* is the wrench in the works that helps create the troublesome intuition—it is the notion that (falsely, but intuitively) ties human design to function by allowing us to think that an item's function is whatever the item was intended for, whether or not the item achieves any degree of match with that intention. This is apparent because the same intuitions are not as strong (and probably thus not as commonly cited) in the realm of organismic traits, where intentions play no role. For instance, although it seems easy to claim that the function of birds' wings is to fly because that is what they were naturally designed for, it is much harder to claim, "The function of those remarkable contraptions (mutant birds' wings that can't fly) is to fly, because that is what they were designed for."[176] And yet, that is precisely the kind of claim that would be necessary to justify a belief that natural design grants function since, if Mother Nature were to be endowed with such a power, she would have to grant functions across the board, regardless of any kind of success or failure (*if* we follow the analogous claims made in the artifact case). My point here is that the very same claim that was being used as a central intuition in the world of artifacts would sound entirely unmotivated in the world of organisms and their traits.

Since intention doesn't play a role in natural design, the question of just what part of the design process grants function here can only come down to the generate-and-test procedure

[175] And this in spite of the fact that the intention of every one of these authors was to argue for biological function!
[176] We would just consider the mutant birds not to have been designed and their handicapped, nonflying wings to be functionless.

(assuming design grants function). But just as we saw with artifacts, this would be circular, since testing (the success or failure of an organism bearing a trait) consists in measuring functionality. The notion of functioning must be independent of generate-and-test design.

There are also counterexamples that will help to dissociate natural design from function. First, there are things that have been naturally designed but that do not function. We call them vestiges. These items may have quite the same form as they did when they were functional; they may even still be capable of functioning if they were to be employed by their bearers, but they have fallen completely out of use because the environmental challenge which their function helps to solve is no longer faced by their bearers. Thus we humans tote around a vermiform appendix, which may be able to aid in the digestion of something we no longer eat, but which goes unused probably our entire lives[177]. Luckily it is not too heavy.

Second, there are things that function but which have not been naturally designed for that function. For instance, the ears and the nose were never designed to hold up eyeglasses, but they function in this way. Likewise the hips hold up pants, and the ring finger functions to provide a location for displaying social cues about mating availability. Although we would not say that these traits have this function, we would say it is a way in which they may function (and TDHF!)

One more issue that bears on whether natural design may grant function is the biological analogue of the design-and-construction gradation problem. We can call it "the developmental gradation problem". Think, for example, of the liver of a goose. We might ask, as the goose develops from embryo to adult, when the liver comes to have its function and by what mechanism. At some point there was a proto-liver, a small mass of undifferentiated stem cells that would eventually become liver cells and we can all agree that this proto-liver neither "had the functions"

[177] This view may be outdated. It is unclear whether our appendix is never used, or only occasionally or non-critically used. All we know for sure is that when the organ is removed, people seem to get on just fine and lead long healthy lives without it. New theories, however, suggest various uses for it (*e.g.*, Bollinger *et al.* 2007; Zahid 2004). Nonetheless, using the traditional view of the organ as a vestigial trait makes for a clear example.

nor served the functions that an adult liver typically does. At some later point in the lifetime of the goose, there comes to be a fully functional liver, and the question is: How and when did the functionless proto-liver come to have the functions of a liver? Just what process conferred a function upon the liver? If it has its function "because it was designed", then the developmental details would seem to be irrelevant and it would have had that function all along, but this is an absurdity similar to that of the metal that will one day become the scissors. If, instead, the function is granted somewhere in the process of development, it would seem arbitrary to draw a line anywhere except the point at which the liver comes to be in a functional relationship—when it comes to play a particular role. But if that is when the liver at last has its function then we should consider function to be not a result of design but of an item's form and the role that that form is able to play (in a particular context), just as we found to be the case with artifacts.

Design and Function

At this point we've doubly dissociated design and function in both organisms and artifacts. The two patterns commonly occur together, but there are regular exceptions. Functioning can occur without any process of design (as in the bottles that can serve as floats or the nose that holds up eyeglasses). And no matter what we take design to be—intention, construction, generate-and-test—it doesn't necessarily imply that its products will be functional . . . except when we take it to be a process that ends only when its products measure up to a functional test. In that case, however, it is function that leads design, not vice versa. As it turns out, this way of looking at it not only explains why functional things are very commonly designed and why designed things very commonly function, but also why this relation is merely a common one and not quite a necessary one.

G. The Function–Accident Distinction

Very likely the central distinction of this analysis is that between the function of something and other things it does which are not its function (or one of its functions) . . . This is sometimes put as the distinction between a function, and something done merely "by accident." Explaining the propriety of this way of speaking—that is, making sense of the function/accident distinction—is . . . perhaps the primary aim of the following analysis.

—Larry Wright (1973)

Making the supposition that functional effects are one kind of effect that a thing may have, philosophers have commonly taken it that if we can find a method for carving up the parent category of *all effects that an item might have* into those that are functional and those that are not, and if that method does not merely pick out a superficial mark of function, then we very likely have at least a close approximation to something that may serve to outline a definitive theory of function. The comparison class that is usually cited is what's called *accidents* or, alternatively, *byproducts*, *incidental effects*, or *mere effects,* and it is thought that a central criterion for a theory of function is that the theory should be able to make what is called the function–accident distinction (Aristotle *Physics* II.5,6; Buller 1998, 1999 [see the epigraph at the start of the current chapter]; Godfrey-Smith 1993; Lewens 2004; Neander 1983; van Parijs 1982; Wright 1973).

We can review the function–accident distinction in terms of our central example, the heart. A heart pumps blood but it also dephosphorylates ATP, creates heat, causes a pulse in the extremities, makes the sounds in the chest cavity that we refer to as "heartbeats"[178], and in some

[178] Of course, as many authors have pointed out (*e.g.*, Wright 1973), once reconstrued in light of a doctor's use of heartbeats to provide medical diagnoses, the heart can be seen to have the function of providing physiological indications, but the difference is important: the heartbeat must be *used by a doctor* in order for the heart to have that

cultures, if it is a chicken heart, may serve as a barbecue item. Not all of these properties are referred to as the heart's function. Some are byproducts—accidental effects of the heart's behavior, or of its form or existence. It seems that only pumping blood is what, in some central sense, the heart is *for*.

Accidental Functions

Some theorists have noticed, however, that there is something of a wrinkle in this function–accident distinction. In particular, there appear to be instances of *accidental functioning*. One of the function statements recalled from earlier can serve as an example.

AF1. (FS14) The bible in the soldier's breast pocket functioned to stop a bullet from entering his heart.

We understand AF1 to be a case in which the bible (or the belt buckle, as the story is sometimes told) just happened to accidentally, fortuitously, be in a place where it was able to intercept a bullet, and still we are comfortable thinking that it functioned to stop the bullet. Boorse and Wright, both of who analyzed this case, also give a few more examples each.

AF2. One squirrel might catch its tail in a crack en route to being run over by a car. (Boorse 1977)

function and, really, we wouldn't say it *has* the function so much as it *serves* the function, until the usage becomes widespread in our culture and its functioning in this way becomes commonplace.

AF3. A bee sting on my nose [might bring] me to a doctor, who spots a curable melanoma on my neck. (Boorse 2002)

AF4. It is merely fortuitous that the nose supports eyeglasses; it is [a] happy chance that the heart throb is diagnostically significant; it would be the merest serendipity if the sixth rib were to be a particularly good pacemaker hook. (Wright 1973)

Though the word "function" was not used explicitly in these examples, the situations in them can easily be described in terms of functioning, despite the fact that each example is based on an accidental occurrence. If a theory of function is meant to distinguish between functions and accidents, then how should we interpret the foregoing situations?

For Wright and for most other theorists that we'll encounter, the conclusion to draw is that all functioning, whether intentional or accidental, fits in a category that is simply *related to* the kind of proper functions that those theorists are interested in. Incidents of accidental functioning are not actually functional, these theorists claim. They are simply random events in which things cause effects that are *similar* to the effects of things that might have real functions. A bible stopping a bullet is similar to a Kevlar vest stopping a bullet, but only the Kevlar vest has that function, and so only the vest's stopping of the bullet counts as *real* functioning. Wright, referring to the common manner of speaking in which we use "function as" and "function to", says, "we signal the difference by a standard sort of 'let's pretend' talk" (1973). He thinks proper functions are real and verb functioning is pretend or illusion or just a manner of speaking[179].

[179] There certainly can be a bit of "let's pretend" going on in these statements. The class of item that a thing functioned as (say, a shield) is certainly not what the thing itself is, and so we *are* pretending the belt buckle or bible is a shield if we claim, for instance that "the bible functioned as a shield". But that's the extent of the make-believe; that is all that the

Boorse (2002) sees it differently. He finds it better to accept accidental functioning[180] as a "weak" sense of function (as in "serving a function") and non-accidental functioning as a "strong" sense (as in "having a function"). Boorse thinks using these two categories of function statements serves him well because it allows his theory to be more broadly applicable (to all phenomena where items function) while also giving him a way to defend against criticisms that his theory doesn't allow for a distinction between functions and accidents. The difference between the weak and strong senses of function, for Boorse, lies in the frequency of their occurrence, with things that have a function being the ones that do function most regularly. I largely agree with Boorse's position here but the issue that he doesn't resolve is the question of where on the spectrum of possible frequencies we might draw the line between functioning and having a function. I suggest that, while the proper-function theorist would need to draw a line (proper functions are black-or-white, discrete phenomena), Boorse does not. We can take just the opposite tack from Wright: Wright claimed to solve the accidental functioning problem by dispensing with *functioning* in favor of *having a function*. Boorse tried to keep both. But if we instead dispense of having a function in favor of functioning, then we need not draw a line at all because functioning is something that really happens in the world, while having a function is just a subjective judgment.

Accidents Without Functions

In the traditional conception of the function–accident distinction, a (proper) function—a class of activity that an item has it in its nature to perform—is being compared to uncontrollable events that occur. This naturally seems like a clear category division along the (admittedly unclear)

"as" signifies. It is important to note where the charade stops. We are not *pretending* that the belt buckle or bible did in fact function. It did. We are not saying that it performed something *like* stopping a bullet, yet not quite. It stopped a bullet. The functioning of the belt buckle or bible was identical to that of a shield; it just happened not to *be* a shield.

[180] Along with other (verb-) functioning, such as the dog's intentional, non-accidental, use of a box as a bed.

lines of "in its nature or not", which then becomes the subject that a theory of function must define (what does the term "in its nature" mean?). But tossing out proper functions in favor of (verb-) functioning, as I've advocated above, simply renders the distinction meaningless since neither functioning nor accidents are "in an item's nature". The former is a relationship to some context that the item plays a role in; the latter is an event that may or may not occur. If we see things this way, there is no problem with the kind of accidental functioning that occurs in bibles and belt buckles that function to block bullets, or in bee stings that function to bring Boorse to a doctor.

The importance of this conclusion should not be overlooked. The function–accident distinction has been used widely to judge whether a theory of function is adequate. But it applies only to a theory that attempts to place proper functions as real properties in the world; if proper functions are illusions, as I've argued, then we are no longer talking about "the nature of" items in and of themselves, and so the function–accident distinction simply has no meaning.

H. Things Don't Have Functions

We are accustomed to hearing about biological functions for various bodily organs. The heart, the kidneys, and the pituitary gland, we are told, have functions—things they are, in this sense supposed to do. *The fact that these organs are supposed to do these things, the fact that they have their functions, is quite independent of what* we *think they are supposed to do. Biologists* discovered *these functions; they didn't invent or assign them.*

—Fred Dretske (1988:91)

This chapter has raised a number of traditional issues that seem to accompany the analysis of function, but it has given very non-traditional answers to most of them. What I have been trying to argue throughout the chapter is that we need to approach the philosophical analysis of function with a "forget everything you know" mindset. In order to make progress, I believe we have to relinquish, all at once, a great many intuitions, preconceptions, and acculturated beliefs that we find coloring the way we currently see function.

Central to that reform will be coming to grips with the idea that things don't have functions (TDHF). We have to throw out the very heart of function analysis—the belief in objective functions themselves—and replace it with a view of functioning as a relation between items and a context of usage. Counter to what Dretske (for one) implies in the epigraph above, proper functions do not exist. But still, function*ing* is an important pattern in our world; it ties together the tools of human artifice—from literal tools such as hammers and pencils to more communal tools such as words, halls of justice, or the scientific method—with the natural products of evolution, and the fact that there is a link that places all of these things on one continuum deserves a good explanation.

A number of other highly reformative perspectives come part and parcel with the disposal of proper functions. First, I think we must relinquish the belief that design—whether by human artifice or natural selection—can grant function. Second, I think malfunctioning items should be seen not as items with functions that fail, but as non-functional items that appear similar to functional ones (an argument to this end appears in the next chapter). Third, I think we need to take care not to be misled when thinking of functional items in terms of the question "What is it for?" And fourth, I argue that the function–accident distinction is an invalid concept by which to measure a theory of functioning. If proper functions don't exist, then accidents can be seen as functional on occasion, and non-functional at other times, rather than being set in strict opposition to functions. If I am right in all of this, then the topic of functioning that I plan to analyze has very little in common with the topic of function that other theorists have previously analyzed.

It may be overstating things to call this "reform" just yet; since I haven't yet produced a replacement for all these notions, at the moment it is more of a demolition—a disassembly of a faulty conceptual edifice in order to recover the bricks, which, in the later chapters, may be useful in building a new structure. But the demolition is not yet complete. A number of other confusing intuitions underlie the various theories of function that have been offered. Let's turn our attention to those theories next.

Chapter V

Functions: Theories

How can we explicate the biological concept of function, and explain what it means to say 'the function of X is Y', without invoking backward causation or panpsychism? Several theories in the philosophy of biology have attempted to solve this problem.

—Buller (2001)

In the previous chapter I discussed a handful of issues that are connected with the ways we conceptualize *function* without yet focusing on any particular theory of it. It will be worthwhile now to examine the many theories that have been offered, not least because the presentation in Part II will lean heavily upon various pieces of those older ideas. Most recent reviews of the literature on function have tended toward a convention of grouping the analyses into two or three main strands, while giving only occasional and dismissive mention to a handful of others that are considered to be more marginal. In this chapter I will review a slightly broader catalogue that focuses on the six categories of theory already listed (on p. 196), while still leaving some work largely unexamined.

The often-overlooked views include those theories that claim that all of teleology arises from the human mind (*e.g.* Ducasse 1925; Nissen 1993, 1997; and Woodfield 1976) or another mind (*e.g.* that of a creator)[181]; they also include the accounts (labeled as "recent pre-history" by Buller, 1999) that really set the stage for the modern debate but which have since been either superseded by or else assimilated into the modern debate (*e.g.* Hempel 1965; Lehman 1965; Nagel 1961; Rosenblueth *et al.*, 1943; Scheffler 1959; Sommerhoff 1950, 1968; Sorabji 1964). The modern debate that

[181] These accounts will be reviewed briefly later.

supplants those contributions is generally taken to have begun about half a century ago, in the early nineteen-seventies.

Of the views that will be looked-over rather than overlooked, the first three—the causal role theory (CR), the selected effects theory (SE), and the replication dispositions theory (RD)—are the usual suspects that, along with hybrid views amongst them, have received the most airtime in recent publications. I'll call them the three *popular* views. The *unpopular* three are what I'll call the goal contributions theory (GC), the valuable effects theory (VE), and the programmed effects theory (PE). I suspect the reason these latter three analyses seldom garner serious attention is because they are each expressed in vague terms. However, I would like the reader to pay particular attention to the unpopular views because, despite their imprecision in terms, I find each of them to be deeply intuitive, and because a slightly richer version of each will play a significant role in outlining the skeleton of the theory I'll later advocate.

I have arranged my review of the literature here with two chief goals in mind. The more important of these goals is to mine useful insights from the existing work. Each theory that has been presented is an attempt to account for some observations and intuitions that a theorist found to be of crucial interest in their analysis and, whether or not the theorist's explanation ultimately succeeds, the observations it is derived from will be informative to my study. A detailed analysis will help to catalog those observations, and the theory presented later can then be held accountable to try to explain the same phenomena as the predecessors it hopes to blend, amend, or displace. That catalog will appear as a series of promissory notes distributed throughout the chapter, all of which I intend to repay in Part II of the dissertation.

The second goal of this review will be to advance a set of critical reflections about each analysis in order to assess each one's suitability to serve as a general theory of both functional and teleological phenomena. I must qualify what I mean by this: The theories we will look at have all

been (claimed to have been) developed for distinct reasons—some have been presented in order to buoy theories of semantics (Millikan 1984); some were developed in order to account for biomedical normality (Boorse 1976); some were meant primarily to analyze and explain complex capacities (Cummins 1983); some were meant to account only for biological function, agnostic of artifact function (Bigelow and Pargetter 1987; Godfrey-Smith 1994; Mayr 1992); and so on. It is often unclear, however, just how narrowly or broadly applicable each author hopes their theory to be since, at times (especially in their critiques of one another), they all lapse into discussion of the more general notion of function or speak more widely of both artifacts and organisms.

In terms of my project, there is no value in critiquing these theorists' presentations with respect to their own stated goals; instead, as I said, I will perform my critique of each with regard to their suitability for sustaining my own explanatory aims. I will proceed bluntly, and in doing so I will undoubtedly step on toes and make judgments that may seem misplaced or unfair, but my reason for doing so is that I am operationally focused only on my own goals: I am digging through other people's answers to their questions in order to find ideas that I can use to answer my questions.

I will follow three central strategies in advancing my diagnostic concerns. The first strategy will be to use a set of examples—I'll call them the "base cases"—to test how each theory accounts for the functional nature of the items in these cases. All three are adapted from the list of function statements in Chapter IV.

BC1. (FS1) The function of a heart is to pump blood in an animal.

BC2. (FS6) The function of a cog is to translate rotational motion in a machine.

BC3. (FS7) The function of a stone paperweight is to hold down papers on a desk.

The first case—the heart—is an organismal trait that has been used as a central example in nearly every analysis of function given in the literature. It could easily have been the function of an eye, a wing, a gill, or a liver; any of these items should be treated similarly by any reasonable theory. But historically the heart has been analyzed most often and so I will keep with that tradition. The second case is a typical part of an artifact—it is a somewhat generic example that tries to capture our intuitions about the ways artifacts are commonly made of interacting small functional parts. The cog should be able to stand in equally for a piston in an engine, a strut in a truss, the screw that acts as a fulcrum in a pair of scissors, or the lid of a disposable coffee cup—any individual part of a multi-part mechanism. The third base case is a less usually called upon artifact example that is meant to be uncontroversial in that we should not doubt its functionality, but that also helps, as we'll see, to highlight theoretical complications in some cases. One class of near equivalents may include simple cups and saucers and other functional artifacts that appear to have no smaller parts[182], yet which serve functions as wholes (*e.g.* holding things); however, found objects that are employed toward various ends (for instance stone doorstops or projectiles, logs used for floating upon, or a stick used to beat the dust from a rug) are near equivalents on another measure because, like the paperweight, they also have no (relevant) evolutionary history or future—an important feature of some of the theories we'll see.

As we go through the chapter, we'll find that none of the popular theories clearly accounts for all three cases, while the unpopular GC and VE theories perform admirably, and the other unpopular account, the PE theory, effectively fails the test. The ways in which a particular theory fails to account for these base cases will, in most instances, suggest notions that will be useful in developing a new theory. It is worth noting, also, that all three base cases are cases of something "having" a function but of course, TDHF! Using these examples allows me to analyze existing

[182] That is, on a certain measure of what counts as a "part".

theories in their own terms, however I will try when I can to reframe these cases as items that are serving functions or being functional.

The second critical strategy I will follow will be to borrow counterexamples from the existing debate. The majority of the work in producing and defending, as well as criticizing, theories of function has been done using conceptual analysis. Because of this tradition, there are thoughtful critics of each theory that have all found useful ways to review each other's work and I will heavily borrow those analyses here in order to avoid reinventing these authors' wheels[183]. As with the first strategy, my intention in using these examples will be to use them to try both to narrow in upon what seems stout about each theory as well as to locate some places in which each may need to be buttressed or reworked (in order to serve my quite general aims).

The third strategy will be to raise theoretical concerns with each analysis (again, with respect to my aims). When employed against the popular theories, this strategy will turn up as efforts to show that there are evaluative-normative (see Chapter I, p. 23, for a definition of this term) or teleological terms implicit in each offering. This suggests that each fundamentally requires underlying concepts of that variety in accounting for functions, even while superficially shunning them, ignoring them, mistaking them, or leaving them as unstated, hidden assumptions. When employed against the unpopular theories, the third strategy will come as an attempt to show that each of these theories, which seem robust against many counterexamples, has a significant vagueness of terms underlying it (which likely accounts for its robustness against counterexamples), leaving the theory unable to do much in the way of real predictive or theoretical work.

In addition, there is one major overarching theoretical concern that can be applied equally to each of the theories: I've just argued in the previous chapter that functions are not the constant sort of properties of objects that we imagine them to be, yet virtually all theorists have been trying to

[183] And to avoid introducing too many more new examples into an already cluttered body of literature.

account for an item's property-functions—its "proper functions"—which I don't think exist. In this regard, I think all six of the upcoming theories are attempts to provide an answer to a flawed question.

A. Causal Roles

We are confronted by the all important question: are those processes in the organism, which we described as purposive, perhaps only purposive in virtue of a given structure or tectonic, of a "machine" in the widest sense, on the basis of which they play their part, being purposive therefore only in the sense in which processes in a machine made by men are purposive; or is there another special kind of teleology in the realm of organic life?

—Hans Driesch (1914)

While teleology seeks to answer a why-is-it-there question by answering a prior what-is-it-for question, functional analysis does not address a why-is-it-there question at all, but a how-does-it-work question. These last are answered by specifying the structure (design) of the system.

—Robert Cummins (2002)

The first of the three popular views that we'll look at is what has come to be known as the Causal Role (CR)[184] analysis of functions. It is generally given in terms of the procedure called "functional analysis" by which an analyst decomposes systems into their functional parts, each of which plays a causal role in the system.[185] The CR analysis considers a function to be *the causal contribution of a part of a system to a behavioral capacity of the whole system* (Cummins 1975, 2002; Davies 2001; Prior 1985; see also Amundson and Lauder 1994; Lehman 1965; Hardcastle 2002; and Hempel

[184] Following a pattern that has become conventional in the functions literature, I may sometimes refer to the causal role theory by the abbreviation "CR". The selected effects theory is also conventionally referred to as the "SE" theory, and I have developed two-letter abbreviations for the other principal function theories, which will show up in due course. Some writers also refer to functions as described by the CR theory as "Cummins functions" for obvious reasons, and I may at times also borrow this name.

[185] This is neither related to the branch of mathematics also called "functional analysis" nor to the psychotherapist's protocol that goes by the same name.

1965). It is important to notice that, on the CR analysis, a function exists relative to an item's context—an item must reside within a particular "system" that has a particular "capacity" in order to have a function.

So the function of the heart is given in the context of the capacity of the whole circulatory system to circulate blood and thereby transport oxygen, nutrients, and waste: its function is to pump blood because that is what the heart does that contributes (causally) to effective circulation. Likewise, the function of a cog is whatever it does to keep the machine that it is a part of running properly (in its overall capacity as a particular kind of machine), and the function of a bird's wing is whatever it does to aid the bird in the capacity of flight. Each of these functional claims is made supposedly in light only of a causal contribution to the capacity of the larger system.

The CR analysis stumbles, however, when asked to give the function of a stone used as a paperweight, doorstop, or projectile, since, in these cases, there appears to be no containing system that the stone is a part of and thus no "overall capacity" that is being contributed to. This stumble suggests that perhaps one of two things is the case.

The first possibility is that our conception of these stones as functioning is mistaken. There are multiple ways that might happen. For instance, perhaps stones used in these ways don't actually function and the idea that they do is an illusion, and so our theory of function just doesn't need to account for them, or perhaps our idea of function divides into two further senses—one that describes parts in systems and another one, for which a separate theory is necessary, that describes items not in systems but which are used whole. At this point we can't rule out such a possibility, but I am unconvinced and I think other considerations that we'll come to later will give us reason to believe we can maintain unity within all perceived functioning.

The second possibility is that the notion of being a "part" of a "containing system" may be just a near approximation of the notion that truly captures or describes all functioning. That is,

perhaps something in the CR analysis as currently stated is imprecise and sharpening it will help find a new rubric under which we can consider parts within containing systems to be akin to non-part items such as the stone paperweight, which are simply used (functionally) by a person. After all, even items such as a telephone (FS9) or a watch or a pocket calculator, each of which are made of many parts but are not themselves parts, still seem to have functions[186] for which they are used and that a theory of functioning should try to account for.

In looking for such an imprecision, one place we might inquire regards the issue of what should count as a system for the CR analysis. And what should count as a part? For instance, do "parts" need to be physically differentiated objects, like a cog or a screw? One might argue that the *lip* of a cup and the *base* of a cup are different parts that serve different functions, but this could also be a slippery slope, as it requires an attempt to draw lines where only blurry boundaries exist (where does the lip of a cup end?)[187]. Does the world come neatly packaged into systems and their parts, or is the human use of these categories due to perceptual or cognitive or cultural biases? We'll look at this problem more, below. Perhaps the category of functional things has substantial overlap with things that we would like to call parts of systems because parts of systems are, for one reason or another, the most *common* type of functional object we see in the world while exceptions, such as stone paperweights, simply go unnoticed most of the time[188].

[186] Or, rather, they seem to be functional (since TDHF!)

[187] Don't the individual masses of each atom in a stone causally contribute to the stone's capacity of being heavy enough to hold down paper? Are those atoms thereby each to be considered functional "parts"? I think perhaps they should, but I don't think that Cummins or other CR analysts usually conceive of parts in this way (see also Wimsatt 1986).

[188] One reason they might be the most common is simply the mathematical consequence that there are more parts of things than there are whole things.

Counterexamples

The stone paperweight is a problematic exclusion—a false negative—for the CR analysis, but the most common criticism of Cummins' statement of function is that it is liberally inclusive of things that don't seem to be functions—that is, it gives false positives. Citing many who have registered similar worries, Boorse (2002, p. 65), puts the point this way:

> It implies that the function of mists is to make rainbows (Bigelow and Pargetter 1987: 184), the function of rocks in a river is to widen the river delta (Kitcher 1993: 390), 'the function of clouds [is] to make rain with which to fill the streams and rivers' (Millikan 1989a: 294), and the function of a piece of dirt stuck in a pipe is to regulate the water flow (Griffiths 1993: 411).

Although each of the so-called functional items here causally contributes to the greater capacity of the larger system it is in, few people would consider any of these items to actually have (or even serve) those functions. Somehow those contributions just don't seem to be functional contributions.

The examples in the previous list describe only physical systems, but Boorse (ibid) adds:

> Moreover, it creates false functions within biology too. Relative to our capacity to die of fluke infestation, our liver's capacity to house liver flukes is its function (Griffiths 1993: 411); relative to our ability to grow gigantic malignant tumors, oncogenes have many functions (Kitcher 1993:390; Melander 1997: 53-4).

These biological examples are perhaps complicated by the presence of other biological agents (the flukes and the cancer cells themselves) which might be construed as alternate and thus independent "systems" with capacities of their own to which a function might be considered relative. Though to simplify this complication one could for instance adjust Griffiths' liver fluke example: Relative to our capacity to die of (or to acquire dementia from) mercury poisoning, the capacity of selenium to react with mercury is its biological function. Also, Matthen (1988, p. 15) adds to the group of biological false positives the fact that, relative to the narwhal's capacity to get stuck in mud, its enormous tusk has the function of weighing down the narwhal.

In anticipation of some counterarguments, Cummins supplemented his definition with a seldom-cited set of conditions that he thought might further define the boundary separating functional things from nonfunctional things. The conditions he gave stated that the extent to which an item has a function, rather than a mere effect, is relative to the degree that the functional capacity is (1) less sophisticated than, and (2) of a different type than the higher-level capacity that is explained; and also relative to (3) the degree of complexity in the system's organization. So he thought, for example, the joint facts that the throbbing of the heart is simpler than and different in type from circulation, and that the circulatory system has a relatively complicated organization, together help us properly attribute the function of the heart as pumping blood, while not attributing such false functions as heart sounds since the latter "differ little if at all in type and sophistication" from the throbbing of the heart that produces them (Cummins 1975)[189]. Paul Sheldon Davies has recently attempted to champion the CR analysis by reconstruing Cummins' three constraints as symptoms of the simpler idea that a function need only play its causal role in the context of a "hierarchical system" (Davies 2001; see also McShea 2012).

[189] It seems, though, that this would also have the bizarre implication that the production of sounds through a process of "throbbing" could never be the function of *anything*, say for instance, an electromagnetic speaker or a wind-up alarm-clock bell. It also seems to imply that if we had a biological organ which was "meant to" create a thumping sound, the CR analysis would have to rule out the production of those sounds just as it does with heart sounds.

Even augmented by Cummins' constraints (or viewed in Davies' terms), though, the CR analysis is difficult to sustain, not least because the constraints all require subjective judgments as to just *how* different in type, *how* much less sophisticated, and *how* complexly organized the relevant parameters are. Opinion on these matters may vary substantially. For instance, it might seem to some but not to others that the narwhal's tusk being heavy is simpler and different in type than the overall capacity for the narwhal to get stuck in mud. And while one might argue that the capacity to get stuck in the mud seems not very complex in organization, the same arguments might erroneously exclude a counterweight from having the function of keeping a window sash from careening back down when lifted, or the lead worn on a free-diver's belt from having the function of drawing the diver deep under water. Why are these various functional weights considered functional but the narwhal's tusk or the avalanche that holds down a skier not functional? Also, using our mercury-poisoning modification of Griffiths' example, it seems that the selenium reaction is simpler and different in type than the overall capacity to die of mercury poisoning, and that the method of death (which first requires the destruction of protective selenoenzymes, and then a set of biochemical processes that result in the neurodegenerative oxidation of now-unprotected brain tissues) is certainly complex in organization[190] but, still, we wouldn't want to thereby consider reactivity with mercury to be the biological function of selenium.

Moreover, if a person had intentionally wedged Griffiths' piece of dirt in the pipe in order to regulate water flow, then it would indeed be functional despite the fact that the human-placed and the accidentally-lodged pieces of dirt will not differ at all in how well they fulfill Cummins' three constraints—the two blocked-pipe systems may be identical except in how and why they have come to be.

[190] However, to repeat, these are indeed subjective judgments, and so it is not entirely clear.

This long and almost indefinitely extensible list of counterexamples shows there may be a missing distinction in Cummins' analysis. We might, however, be able to mine the examples for commonalities that could be used to strengthen the model. One thing of note is that the counterexamples seem to be situations either in which effects of an item contribute to the occurrence of some event regardless of whether that event is desirable or not (rainmaking clouds and rainbow-making mists), or in which the biological goals of survival and reproduction seem to be thwarted by a capacity that an item contributes to (oncogenes, mercury, liver flukes, or the narwhal's tusk). We could add a third category, similar to the second, in which the intentional goals of an actor are thwarted by a capacity that an item contributes to. For instance, a software bug—a bit of delinquent code—which is a part of a system (the program it occurs in or the hardware that that program controls) and which regularly causes failures in the behavior of the machine—say, a missile guidance system that fails to track its target due to an accumulated floating-point error in its timing clock[191]. Such a bug seems to fit Cummins' constraints of being simpler and different in type from the tracking failure it contributes to, and the overall system is certainly complex in organization. One can imagine any number of other situations in this third category, in which a part of a poorly designed or poorly built machine or system does not contribute to what one intends for the machine to do when using it. Together, the three categories (physical processes that just happen whether or not agents are involved and things that are counter to either biological goals or psychological agents' goals) are the opposite of another concisely definable (though still not yet precise) category: things that serve agents' goals. This is the intuition that leads to the goal contribution (GC) analysis discussed later in this chapter.

[191] This example is based on a real bug that existed in 1991 in the Patriot Missile systems manufactured by the Raytheon Company, and that contributed to a well-documented catastrophic failure.

Smuggling

It is important to note that the CR analysis as it is standardly presented is an *anti*-teleological view. Cummins first offered his version of the analysis as a response to earlier attempts at defining functions in terms that could explain a trait or an item's presence in an organism or other system (Hempel 1959; Nagel 1961), and he protested that this tendency of those earlier theorists was a "failure to distinguish teleological explanation from functional explanation" (1975). I think Cummins' distinction here is a false one: My claim, later, will be that there is no functional explanation that is not ultimately also teleological explanation, and the CR analyst's presumption that there might be is a mistake. Examining the method by which Cummins attempts to excise teleology from his account will help us develop the suspicion that this method is somewhat disingenuous.

At its core, the CR analysis appeals to a higher-level "capacity" of a containing "system". Both of these words, "system" and "capacity", were chosen carefully because they appear to be teleologically innocent terms. They imply that, as analysts, we are looking at well-defined objective categories in the world called systems and the objective processes that they undergo, the results of which are called capacities. Cummins would have us believe that, in functional analysis, we are examining *what* a thing is able to do and *how* it does it without regard to *why*.

While in their most generic forms, the terms "system" and "capacity" do seem to discard any subjective element that might invite teleological thinking, it is the specific choice made by an analyst of each particular system and of its capacity, during the individual analysis of any function that smuggles the teleology back in (much like the choice, made by a dieter who eats only salads, to smuggle the calories back in each time with a rich and creamy dressing). In particular, if we choose an inherently teleological system as the reference context when analyzing a function then, despite

appearances, the CR analysis is not as teleologically innocuous as advertised[192]. Let's have a look at our examples again.

For starters, Cummins (1975) reminds us that the function of chlorophyll is in contributing to the photosynthetic production of sugars; the function of nerve tissue in animals is in aiding coordinated activity; and the function of a heart in vertebrates is its contribution to the circulation of the blood. In all of these, and any other, biological cases that the CR analysis seems to correctly classify as functions, the capacity to which an item contributes is not simply a *physical process*; it is not something that just happens to happen in the world. Each such capacity—producing sugars, coordinating activity and circulating blood—also ultimately if not directly plays a role in assisting an organism to survive and reproduce, which I claim is an archetypal teleological capacity.[193]

It is clear that this last claim remains partially unsubstantiated without an underlying, objective, definition of what a teleological capacity is, and many may be apprehensive about whether such a definition can be given (I will make that argument later). But even if we are skeptical about the existence of objective teleology, still we can notice the categorical differences between those systems and capacities that are required to make Cummins' offering work and those systems and capacities that, as counterexamples have shown, cause it to clash with our intuitions. Let's continue to develop those categories.

On the one hand, if we group the CR analysis' true positives[194] (hearts, chlorophyll, nerve tissue, cogs in machinery, and even a clump of dirt intentionally placed in a pipe to slow water flow) along with its false negatives (stone paperweights and doorstops), we find that each example has a goal or purpose *implicit* in the "system" in relation to which the function is defined—there is always,

[192] Amundson and Lauder, in their (1994) paper "Function without Purpose", buy it. They argue for the aptness of the causal role analysis in biological practice and, as their title suggests, they see Cummins as having "introduced a novel concept of function in which the specification of a real, objective goal simply dropped out." I think they've had the wool pulled over their eyes. Perhaps, yes, the *specification* of a goal has dropped out, but the *unspecified* goals are implicitly smuggled back in.

[193] However, this is shorthand for what each ultimately contributes to.

[194] Also called "hits", in signal detection theory.

ultimately, an artifact that serves a user's goals or an organism whose traits and behaviors contribute to its overall goals of survival and reproduction. The systems that happen to allow the CR analysis to succeed can all be seen as being somehow goal-directed at the system level.

On the other hand, consider both the false positives given by the CR analysis (selenium's reaction with mercury, mists that make rainbows, clouds that make rain, a bug in the software, a wrench in the gears,[195] and a clump of dirt accidentally lodged in a pipe) and its true negatives[196] (none have been mentioned yet but let's say, for example, a rock or a puddle not being used for anything). All are, simply, physical objects or processes, only a subset of which—the false positives—are parts involved in what can be called systems. And none of these have any necessary relation to goal-directed artifact usage or organisms' biological goals.

The division is clear. If, in practice, the CR analysis only ever uses artifacts and organisms to provide the relative context to successfully sort functioning items from non-functioning ones, and if it also fails to sort them correctly when it does not use these contexts, then CR analysts must admit this consistent bias as a tacit, hidden constraint of their theory. In short, the causal role analysis is nearly right when it says that a function is a causal contribution to a capacity. What it fails to state is simply that it is not a contribution to just any capacity of any system that counts, but, in particular, a contribution to a goal-directed capacity of a goal-directed system.

[195] One might wonder whether there is goal-directedness in "the software" or "the gears" in these two examples, since they are causally involved in artifacts that serve intentional purposes. However, on the CR analysis, one would then be attempting to say: "the function of the bug (or the wrench) is its causal contribution, relative to the capacity of a missile guidance system to *miss* its target. Missing a target is not in fact a goal, though; it is a failure. And so no goal of the artifact or its creator or user is thereby inherited into the CR analysis.

[196] Also called "correct rejections", in signal detection theory.

Systems and Their Parts

One attempt at accounting for the stone paperweight might consist of including the paper and the stone together as a system. Doing so seems like an arbitrary move, though, since the CR analyst doesn't make the same move for most other artifact functions. It seems strange to include the paper (the thing being held down) as a "part" in the "system" that holds down paper[197]. We don't include the road as part of the car when assigning a function to tires. Still, this consideration helps raise the question: If functions are relative to an item's role as a part in the context of a system, then what should we count as a CR "system" or as a "part" of one?[198]

Certainly, if the CR analysis is to be an objective theory of function (as its proponents usually claim it to be), one cannot merely claim that any arbitrarily defined grouping of parts and processes is "a system" in order to justify attributions made by the theory. There has to be rhyme and reason to the selection process. And yet . . . what natural ways do we have of bounding systems?

Although physicists, especially thermodynamicists, do prolifically use the term "system" (thus speaking of "open", "closed", and "isolated" systems), so far there are no physically objective criteria by which these definitions can be made precise. There is always bleed-through, at the edges, where the idealized boundaries of such systems fail. All physical systems in the universe are, ultimately, open systems[199]. Idealizations are often practical expedients that allow the fields of physics and thermodynamics to proceed *usefully*, but the (good) scientists involved in such practices

[197] Another possibility would be to include both the stone and the person placing it upon the paper, together, as a system. I'm not opposed to the idea of involving the user of an artifact in fixing the artifact's function, but to do so would directly invite conscious goal-directedness into our analysis of artifacts—a move that the CR analyst would probably prefer to avoid.
[198] See also Millikan (1999) for a different yet still critical analysis of the same question.
[199] A perfect thermos is as physically impossible as a perpetual motion machine.

are keenly aware of their assumptions and of the fact that all their models have limits to their generalizability.

So without physics to lean on, where do CR analysts draw the bounds for their systems? As I suggested in the previous section, it seems they are drawn arbitrarily or subjectively by the analyst, wherever it seems intuitive (or, perhaps worse, convenient) for the purposes of their analysis; it seems that goal-directed systems or systems held in a causal-contributory relationship to a goal-directed system are always chosen. But it also seems that while the chosen systems appear to have this goal-directed nature, they are not in fact defined in any other precise terms whatsoever from which a general definition for "system" could be derived. This sounds like a broad claim to substantiate, but while I will review a few examples here, generally I leave it up to any CR analysts who wish to sustain their view to prove this wrong[200].

Let's take the example of a common electromagnetic speaker. Where are the boundaries of this device considered as a "system", if we are to use that system to grant functions to its parts in relation to it? What counts as a "part" of such a device? To a first approximation, most people would probably include the fixed ferromagnetic ring and electromagnetic copper wrappings near it, as well as the cone-shaped diaphragm or membrane attached to the electromagnet, and the solid frame that holds all the rest of the parts in their places. A slightly more detailed review might note that there are some screws and adhesives involved as well as some connecting wires and so on. A CR analyst very likely would call this entire assembly the analyzed system, and its systemic capacity, to which all those parts contribute, would obviously be to produce sound[201]. But is this characterization sufficient? Should we not include the air both within and surrounding the speaker,

[200] Again, CR analysts may not care to sustain their view for the general purposes that I have in mind; they may choose to limit their view to a subset of functioning for some other reason. And that response is just fine as long as such theorists explicitly avow the corresponding limits of the scope of their theory.

[201] Let's ignore the fact that *sound* itself is a complicated phenomenon involving the sensation and perception of a cognitive agent, and instead just make the simplifying assumption that sounds are longitudinal compression waves in a gas.

without which the system will fail to produce sound waves? Might we include the speaker-wires that carry power and a signal to the magnetic assemblage? And, if so, mightn't we also consider the signal-production system, whatever it may be . . . and the power production system, whatever that may be? Maybe not. It seems that including a television or music player and a hydroelectric power plant in our definition of a speaker goes too far. But that is an arbitrary decision. There is no *principle* by which we can easily generalize such a judgment to all cases of "systems". We bound the system where we do for other, intuitive reasons that are not explained in the CR analysis. If such unstated intuitions play a role in our analysis then, as was mentioned in the previous section, it is disingenuous to state the process of analysis without explicit mention of them.

Cummins seems to be aware of the system-bounding problem, though, for reasons unclear to me, it doesn't seem to concern him greatly.

> Indeed, what makes something part of, *e.g.*, the nervous system is that its capacities figure in an analysis of the capacity to respond to external stimuli, coordinate movement, *etc.* Thus, there is no question that the glial cells are part of the brain, but there is some question as to whether they are part of the nervous system or merely auxiliary to it" (Cummins 1975, footnote 18, p. 761)

If we take this at face value, if we take it that what makes something part of an analyzed system is that the part's capacities "figure in an analysis of" the system's capacity, then we should in fact include, as parts of the electromagnetic speaker, the power plant, the music player, and the air—all of which absolutely "figure in an analysis of the capacity" to make sounds.[202]

[202] Not to mention that if Cummins' proposal for defining a system is based in choosing the parts that figure in an analysis of it, then his analysis of the functions of parts in terms of a system becomes circular: A function is the role an item plays in a system that is made up of parts that play a role.

This is not an isolated or unusual case, either. The same question applies to any artifact. Televisions or computers clearly use electric power in the same way the speaker does. Automobiles are fueled by a complex economy of oil drilling and distribution. Items such as flush toilets and showers require pressurized water feed pipes and drains as well as pumps that may be operating miles away, in order for their primary capacities to operate. Do we include the sun as "part of the system" when analyzing a solar-powered calculator? Do we include it when analyzing a solar oven? Do we include the river, when analyzing the capacity of a watermill to grind grain? How much of it? And, in this case, should we also include the sun, the energy from which carried the water to the clouds that dumped it into the watershed of that river?

Cummins' example of glial cells opens the door to biological cases too. Where shall we bound the circulatory system or the pulmonary system or the nervous system in an organism? Are glial cells "part of the nervous system or merely auxiliary to it"? Do we or don't we include the innervation of the heart in the circulatory system? If so, how far back do we cut those nerves when carving out the circulatory system? Is the brain stem far enough? Why? What principle can we possibly use for bounding our system anywhere if we are interested in lower-level capacities "that figure in an analysis of" a systemic capacity?

The issue here can perhaps be cast in terms of the more general philosophical problem of open-ended causation. Since every event in the world is caused by previous events and those events are themselves caused by even further previous events, then when a theory gives an answer in terms of causation, it invites an unending series of answers spanning the period from the event in question all the way back to the beginning of time. Likewise, when one asks what future events are caused by a particular event, there is an ever widening light-cone of future causal consequences none of which is *the* effect, but each of which is *an* effect, of the event in question. Because of this, defining a

system in terms of the items whose causal capacities figure in an analysis of the capacity of the system fares no better than defining a system in physical terms (or subjective or intuitive terms).

This may sound like a negative conclusion, but I don't mean to suggest that it is hopeless to try to define a system. What's more, as it turns out, I think that, in the absence of physical criteria, causality *is* a good starting point for a definition of system, though it needs to be used in a less simplistic way. I claim that the presence of certain causal organizations can be an antidote to the pathological open-endedness of causality. A certain kind of causal bound can create a form of containment that physical bounds never could. This antidote will be an important piece in making the theory I'll later advocate work.

Functions Are Causal Roles

Though I've just expressed concern with the notion of causation, I also think that the focus on causation is the most essential kernel of truth in the CR analysis. A majority of functioning items—even the stone paperweight—do in fact appear to causally contribute to *something*, even if that something is not always simply a "capacity of a containing system". The pumping of the heart plays a causal role that contributes to circulation, the turning of a cog plays a causal role that contributes to the operation of the machine it is in, and the stone paperweight plays a causal role that contributes to its holding down papers (and our wanting it to hold down papers). And so, while the "system" and "part" portions of the CR analysis seem to need some kind of adjustment, the "causation" portion, though still quite thorny, seems nonetheless highly relevant to the notion of function.

In fact, as we go through the remainder of the theories that comprise the debate, we'll find also that each one takes almost for granted that a function is a kind of causal contribution. Three

theories (SE, PE, and VE) are given in terms of "effects"—those effects that a trait or an item (of a type described by the particular theory) has; of the other two theories, one (GC) is given directly in terms of "causal contributions" to goals (Boorse 2002), while the other (RD) is given in terms of the disposition of an item to play a particular (causal) role (Bigelow and Pargetter 1987:196). Everyone seems to agree that there is something fundamentally central in the concept of function about "playing a causal role". It remains to be seen what we can make of this insight, since the notion of causation raises its own philosophical problems—problems that have not yet been deeply examined in the functions debate. As I said, I think the worst of them can be evaded, but we'll have to pick that thread up again later.

Other features of the CR analysis include, first, the fact that it respects a notion of *hierarchy*[203] (even if it is not clear why or how); second, that it finds functions to be relative to a context, which is inconsistent with the idea of a "proper" function (even as the analysis still seems on the surface to be about "the function of"); and, third, it respects our understanding of functional equivalence (the idea that replacing a part in a machine—or organism—with another that plays the same causal role still allows the machine or organism to operate properly)[204].

In addition, there are two issues that I found unconvincing about the CR analysis but that raise important questions that need to be addressed by any proposed amendment of the analysis. First, I've emphasized my disagreement with the anti-teleological stance of the CR analysis, though I admit that an alternative or updated offering would still need to be consistent with the *reasons* why Cummins and other CR proponents provided their anti-teleological offerings—namely, that any theory of function needs to be consistent with the modern scientist's view of materialism.

[203] McShea also put his recent theory of ("seeming") teleology (though not specifically of function) in terms of hierarchical organization, though his is a somewhat more physical sense of hierarchy, given in terms of compositional containment, rather than a causal sense of hierarchy.
[204] Other theories do this too, but only because they are couched in terms that ultimately refer to a causal contribution.

Second, while a saucer or a stone paperweight reminds us that not all items that function have a containing system with a superordinate capacity, still a structure of that sort seems to be extremely common—so common that the exceptions appear to CR theorists to be systematically ignorable. A new theory that intends to make sense of the phenomena of functions and the causal roles played in them should at least make provisions both for those situations in which a functioning item does contribute to the larger capacity of a containing system and for those situations in which it does not, and it should attempt to describe some relationship between the two categories.

B. Selected Effects

Items have functions when their being there depends on reproduction from ancestors having similar traits, these traits having been causally efficacious in helping to produce these items, and these traits having been selected at some point in this history for their capacity to make this kind of contribution.

—Ruth Millikan (1993)

Our second popular view is called the Selected Effects (SE) theory, or the historical or etiological[205] theory of functions. Allen and Bekoff (1995a) have given it the moniker "The Standard Line", as there appear to be many more defenders of versions of this view than any other (*e.g.*, Adams 1979; Ayala 1970; Brandon 1981; Dennett 2014; Enç 1979; Garson 2012; Godfrey-Smith 1994; Griffiths 1993; Mace 1935; Millikan 1984, 1989a, 1989b, 2002; Mitchell 1995; Neander 1983, 1995a, 1995b, 1998; Williams 1966; Wright 1973; not to mention the abundance of plural and so-called unifying theories that are based largely on the SE theory[206]). Allen and Bekoff (1995a) put it this way: "A trait's function or functions causally explain the existence or maintenance of that trait in a given population via the mechanism of natural selection". A bit more pithily, an item's functions are *the effects of the item for which it was selected by natural selection.*

Typically, the SE analysis is expressed first in terms of its original formulation by Larry Wright, and then modified by its now more "standard" and, it is sometimes claimed, more robust formulation, given by a series of authors beginning with Karen Neander and Ruth Millikan. I will follow in that tradition.

[205] The word "etiology" (and its adjectival form, "etiological") is commonly used in the literature on functions. It simply refers to the historical reasons for a thing coming to be the way it is. I will try to stick to the terminology of "Selected Effects".
[206] These will be described briefly later.

Wright

Wright's oft-cited and now-classic formula is the following.

> The function of X is Z means
>
> (a) X is there because it does Z,
>
> (b) Z is a consequence (or result) of X's being there. (Wright 1973)

Although Wright leaves it unmentioned, to understand him properly we need to interpret the shorthand found in some of his words in two very different ways depending on whether we are talking about organisms and their traits, or artifacts and their parts. The ambiguous terms are the word "because" from part (a), the word "consequence" from part (b) and the word "there", which appears twice, once in part (a) and once in part (b).

In the case of organisms, these words bring to mind natural selection as our explanatory mechanism: X *exists* because it does Z and doing Z is *what X's ancestors did successfully in order to replicate, and thus cause X to exist*. In the case of artifacts, these words should be interpreted in terms of human intentions: X *either exists or is located where it is or both* because it does Z and *somebody knows X does Z, and wants it to do Z, and so they built X and/or put X in that location* in order for it to do Z[207].

So, for example, "the function of the heart is pumping blood" means, (a) "the heart is there because it pumps blood", and (b) "pumping blood is a consequence of the heart's being there", wherein (a) is to be interpreted to mean hearts as a category are there (partially, but importantly) because all members of the class of their ancestors' hearts pumped blood successfully, thus contributing to the existence of current hearts. This version of Wright's view is a theory primarily

[207] Wright (1976) clarifies the term "is there" from part (a) of his schema, suggesting that it should best be interpreted to mean, "came to be there". My interpretations in this paragraph are consistent with that.

about types, though one could (and I think Wright intends to) easily extend it to tokens by the simple logical move that a token heart, for instance, inherits most of the properties of the heart type.

For artifacts, however, it is a purely token theory. "The function of the cog is translating rotational motion" means (a) "the cog is there because it translates rotational motion" and (b) "translating rotational motion is a consequence of the cog being there", wherein (a) is to be interpreted as "the cog is there because a person who knows it will translate rotational motion has willingly put it there for just that reason." Similarly, we can claim that "the function of the stone paperweight is holding down paper" means (a) "the stone is there because it holds down paper" and (b) "holding down paper is a consequence of the stone being there," because we know that a person placed the stone paperweight on the paper with the intention of it holding down the paper (though of course this is not why the stone *exists*).

Is Wright Right?

As we just saw, Wright's view seems to describe functions in both of the main categories of organism and artifact successfully, but it does so by relying on an ambiguity in some of its terms that disguises an unarticulated difference between the two applications. When the difference is made explicit, though, the theory divides in two—one theory for each category[208]. As was pointed out by Boorse (1976), Wright's failure to make this division explicit also leaves the schema vulnerable to two categories of counterexample. The first consists of artifacts that are there (partly) because of selection—because of what they do that contributes to their own existence.

[208] Also, if we were to try to explain the functions of *behaviors*, such as those in FS10 and FS11, we would need to put a third, more explicit disjunct into the definition since "is there" does not apply to behaviors. Wright's formula would then look something like "X is there (or is performed) because it does Z and Z is a consequence of X's being there (or X's being performed)". But this doesn't have the kind of circular causality in it that "being there because of what it does" creates, and it isn't clear how to remedy that. Inserting FS10 into the formula, for instance, would produce: "Putting a stamp on a letter is performed because it ensures that the letter gets delivered and ensuring that a letter gets delivered is a consequence of putting a stamp on the letter" . . . which sounds more tautological than teleological.

> Suppose that a scientist builds a laser which is connected by a rubber hose to a source of gaseous chlorine. After turning on the machine he notices a break in the hose, but before he can correct it he inhales the escaping gas and falls unconscious. (Boorse 1976)

In this example, Wright's analysis would have us assign releasing gas as the function of the break in the hose: (a) the break is there because (among other things) it releases gas (thereby preventing anyone from repairing it) and (b) releasing gas is a consequence of the break being there.

Bedau borrows another similar example (but not its analysis, he notes) from Robert van Gulick.

> Consider a stick floating down a stream that brushes against a rock and comes to be pinned there by the backwash it creates. The stick is creating the backwash because of a number of considerations, including the flow of the water, the shape and mass of the stick, *etc.*, but part of the explanation of why it creates the backwash is that the stick is pinned in a certain way on the rock by the water. (Bedau 1992)

Bedau asks pointedly, "Given that the heart example and the stick example involve a similar sort of etiology [meaning that the reason the heart is there is analogous to the reason the stick is there], why is only the heart teleological?" This is a good question.

Boorse's second category of counterexample involves organismic traits that are there because of human intentions (the way, say, the cog or paperweight is). He illustrates:

> A man who is irritated with a barking dog kicks it, breaking one leg, with the intention of causing the animal pain. The dog's pain is a result of the fracture, and the fracture is there because its creator intends it to have that result. (Boorse 1976)[209]

Of course, we wouldn't be willing to assign to the fracture the biological function of causing the dog pain. Even if we were to distort our concept of function in that way, it would certainly not be meant in the same sense as we mean when we say the heart has the function of pumping. The fracture could only be functional in the sense that it is an artifact—it might serve a function for the man, but not for the dog.

Another way Wright's schema assigns functions where it shouldn't is in the case of vestiges. When the environment changes suddenly in such a way that a previously functional trait no longer does anything, Wright would still consider the effects of such traits to be a consequence of the trait being there and the trait to be there because ancestral traits of the same kind were selected for that effect.

McLaughlin (2001) points out also that Wright's schema grants functions to whole organisms (and not just their parts), though most of us would not. For instance, McLaughlin notes, elephants are there because they replicate themselves and elephants' replicating themselves is a consequence of their being there.

So, yes, there is a sense in which Wright has accounted for all three of our base cases—each fits his formula, naïvely interpreted. However, after noticing the ambiguity of his formula and of its applicability across the cases, we must admit that he has not in fact given us a singular theory that

[209] I have found that some readers misinterpret Boorse's example. I will try to explain further: The (cruel) man in the example is meant to have fully intended the broken leg that causes pain. He is not just meant to have intended to cause pain with a kick that also accidentally breaks the leg. Boorse's first sentence makes this unclear by being specific (but not specific enough) about an intention. The second sentence clarifies, but it is easy to miss, especially if we, in our hearts, want the man not to be quite that cruel. The claim is: "the fracture is there because its creator intends it . . ."

accounts for all the base cases, nor one that correctly distinguishes other functions in the world from non-functions, such as in the above counterexamples. The implicit organismal and artifactual halves of his theory have little in common. Nonetheless, there is something intuitively teleological about an item causing itself, and I intend to keep sight of this idea for later.

Evolution

Later authors evolved Wright's formula into its more modern form, described earlier. To repeat, an item's functions are the effects of the item for which it was selected by natural selection. Some of these authors (*e.g.* Neander 1991, Godfrey-Smith 1994) seem interested only in the functions of organismic traits, leaving artifacts and behaviors aside and embracing just the biological half of Wright's formula. Most intend to account for artifacts in one fashion or another. For instance, Griffiths (1993) attempts to wrap artifacts in the cloth of natural selection by focusing on the way many artifact designs are evolved by human selection, with later generations of tools being based upon the more successful (*i.e.*, useful) ones from earlier generations. Kitcher (1993), similarly, attempts to unite organisms and artifacts along the lines that both have been subjected to a process of design. Millikan makes a somewhat more sophisticated move in her case for proper functions, and since her theory raises many issues that will be important to developing our later theory, we will turn our attention to it now.

As was mentioned earlier, Millikan uses the term "proper function" to emphasize that the functions she is interested in are *properties* that, once established, inhere in or belong to the objects that we describe with them.[210] And as you may recall, I am not enthusiastic about this idea; let us,

[210] Neander (1980, 1983, 1991) gave an earlier, very similar, version of the theory of proper functions; however, because she focused only on biological functions, her theory is not quite as complete as Millikan's. Here is Neander's formula for the proper functions of biological traits: "It is the/a proper function of an item (X) of an organism (O) to do that

however, set that criticism aside for the moment and look at the form of Millikan's offering. "Proper *bio*functions" or "*direct* proper functions", Millikan says, get their status by being members of "reproductively established families", which she explains are, more or less, the results of natural selection[211]. She then introduces the notion of "*derived* proper functions"[212] for items that, on their own, are not members of reproductively established families, but that nonetheless derive their functional status from being *produced by* an item that has a direct proper function, *and* having (ancestrally) contributed to the selection of the item that produced them.

In Millikan's own words:

> Putting things very roughly, for an item A to have a function F as a "proper function", it is necessary (and close to sufficient) that one of these two conditions should hold. (1) A originated as a "reproduction" (to give one example, as a copy, or a copy of a copy) of some prior item or items that, due in part to possession of the properties reproduced, have actually performed F in the past, and A exists because (causally historically because) of this or these performances. (2) A originated as the product of some prior device that, given its circumstances, had performance of F as a proper function and that, under those circumstances, normally causes F to be performed by means of producing an item like A. Items that fall under condition (2) have "derived proper functions", functions derived from the functions of the devices that produce them. (Millikan 1989, p. 288)

---

which items of X's type did to contribute to the inclusive fitness of O's ancestors, and which caused the genotype, of which X is the phenotypic expression, to be selected by natural selection." (Neander 1991)

[211] Actually, Millikan's theory is a fair bit more nuanced. But for our purposes here, we needn't focus on too many of the details. Her work was meant to support a more complicated analysis of biosemantics that goes well beyond the central topics of this thesis.

[212] As well as "relational proper functions" and "adapted proper functions"—two notions that address aspects of environmental relativity and plastic responsiveness in proper functions; but, once again, not all of Millikan's details are relevant to our analysis.

For Millikan, then, an organism's genes are copies of prior items, and so they have direct proper functions according to her condition (1), but the heart is a "product" of those genes, and so it has a derived proper function according to her condition (2). An artifact, such as a cog or a stone paperweight, falls under condition (2), thereby allowing her theory to neatly account for all three of my base cases. In fact, for Millikan, artifacts of all sorts have derived proper functions simply by being produced by members of reproductively established families, and she cleverly extends the notion also to behaviors performed by members of reproductively established families (see also Millikan 1999).

This formulation differs from Wright's in at least two interesting ways. First, since Millikan requires functional items either to be members of reproductively established families or to be a special kind of product of those members, as outlined above, she cleanly avoids the counterexamples presented against Wright. Boorse's chlorine leak and Bedau's stick in the stream, for instance, are neither. Second, Millikan grants artifacts their functions not via a wholly distinct schema, but via a branch of the same schema that allows her to grant them to organisms and their parts: at the root of each there is membership in one or another reproductively established family.

What is particularly interesting to me about Millikan's construction is this relationship between *direct* and *derived* functioning. The notion that an item can somehow be derivatively functional is an important one, which accounts for the strange gap in animacy between the two major categories of functional items (living organisms and inanimate artifacts) but one that I don't think Millikan explored quite far enough. I will employ a modified version of it later.

Purported Counterexamples

The more modern SE analysis has been subjected to its share of counterexamples also (some of which may also apply to Wright's earlier version). The three most commonly mentioned are those of clay crystals, the immune system, and various types of selfish DNA. I'll introduce each one here. In the end, however, I find each of these examples to be largely inconclusive and thus not particularly damning for either Millikan or Wright.

Clay is composed of countless tiny crystals (in a mineral class called phyllosilicates) and it has been discovered that these crystals may have random variations—imperfections in their crystalline structure—that are subject to heritability through the process by which crystals grow and cleave. As it turns out, some of these imperfections are able to influence the likelihood of their own proliferation in a population of reproducing crystals (Cairns-Smith 1982, 1985; see also Bedau 1991; Dawkins 1986). Because of these facts, clay crystals are subject to a standard form of natural selection: they have reproduction, heritable variation, and fitness. Despite this, the intuition that most function theorists have is that the imperfections in these crystals have no purpose and no function (*e.g.* Bedau 1991; Boorse 2002; Lewens 2004; Melander 1997). Thus the behavior of these imperfections is often cited as evidence that having a selected effect is not sufficient for, and possibly not at all constitutive of, having a function (although, of course, TDHF).

I am not convinced. I think the SE analysis may be able to slip past this counterexample without sustaining significant damage. I think the logic is sound, but what I question is the intuition that the variations in these crystals are necessarily functionless. This intuition seems based firstly in the assumptions that such crystals appear neither to be alive nor to be the kind of thing that could be a beneficiary, and secondly, in an unwillingness to label as purposeful something that is explainable by physics alone. Bedau (1991) for instance, states a claim of this sort explicitly, in favor

of his intuition that clay crystals do not have a function: he claims that crystals are simply not animated, living beings; they cannot benefit from the service of functional activity. I must admit that I have the very same gut instinct. But in a world where everything is explainable by physics alone (albeit not always usefully or concisely) and yet where purposeful things also exist, there must be a boundary, somewhere, where purposive, minimally lifelike behavior begins to appear in phenomena still physically explainable—a boundary where both levels of explanation may be useful or applicable or, at the very least, comprehensible. Seeming biological in some ways and yet not in others, these crystals lie close to that boundary and, it seems to me, we cannot know whether they are on the interesting, minimally purposeful side or the almost-but-not-quite purposeful side unless we have a productive and detailed theory of vitality or purposiveness or, for Bedau, of benefit or value. At the moment, no theorist who has cited intuitions about clay crystal variations being functionless has given such a theory. I'll examine the same case again later, but for now we can say at a minimum that the heritable variations in clay crystals are not a clean and clear, central counterexample to the SE analysis[213]. Millikan makes much the same judgment: Of the claim that the SE theory would be forced to assign a function to these crystals, she says, "that is fine by me", since it similarly allows the SE theory to assign functions to, for instance, "learned behaviors, artifacts, words, [and] customs." (1993:39).

The immune system (of humans and other vertebrates) is another type of system in which a form of natural selection takes place on a non-genetic entity, which has prompted some to argue that there is selection without function (Matthen and Levy 1984; Matthen 1997). Antibodies, which are present in the body in vast variation, are selected by their ability to "match" with antigens, an event that determines their rate of reproduction, thus contributing to a system of heritable variation, reproduction, and fitness that results in differential reproduction within a population of

[213] Two similar but less natural counterexamples are the example of ball bearing cloning offered by Schaffner (1993:383-4) and the example of artificial selection by Naziesque mad scientists devised by Plantinga (1993:203).

antibodies[214]. Matthen (1997) argues that isolating this system of antibodies and their selection process from a mammalian body—say, extracting the components of the immune system from the blood and marrow where they normally reside, and installing them in a jar or test tube—would not change the antibodies, nor the procedure of selection they might undergo, yet it would render them non-functional, because they would no longer serve a body immunologically. As with clay crystals, the immune-system thought experiment is meant to dissociate having been selected for an effect from having a function.

My concern with this proposed counterexample is similar to the one I expressed concerning Bedau's intuition about the clay crystals, and similar also to the concern I hinted at with the examples of oncogenes and liver flukes in an earlier passage.[215] It seems to me that underlying many intuitions about when a thing does or does not have a function there exists a prejudice about whole organisms being the only possible recipients of the functional service. Matthen would not see antibodies *in vitro* as being functional because their behavior does not contribute to an organism. But it seems to me that we would need a theory of what types of patterns can be beneficiaries (or, at least, recipients of functional contributions) before we can make judgments of this kind. There is a sense—even if a weak one—in which antibodies themselves are patterns that might be seen as beneficiaries that could be served by their own (possibly naturally selected) functioning. It is the same sense in which oncogenes and liver flukes can also be beneficiaries served by their own functioning despite their injurious effects upon their hosts, and the same sense in which selfish DNA (a more convincing example, which we'll look at shortly) can benefit from its own functioning. Again, the central point here is that while Matthen's example of the immune system

[214] Since the variation seems to be prior to, rather than a result of, reproduction, this form of natural selection is not the paradigmatic form in which change accumulates over the course of many generations. But let's leave that concern aside.
[215] I will leave aside the concern that the segregation of immune components from the rest of the body in Matthen's proposed counterexample may be an impossible task.

certainly does raise some doubts, it doesn't seem to be a clear, central counterexample against the SE theory.

The example of the liver functioning to house flukes brings to light a fact of parasitism worth mentioning: an item such as a liver may function both for an original owner and for a parasite although *differently so*. The beneficiary of an item's functioning need not be the entity that we traditionally describe as the owner of an item. A hermit crab benefits from the functioning of a gastropod's shell, and every non-photosynthetic creature makes energetic functional use of another organism's tissues or products. Here's a fascinating case: the tongue-eating louse (*Cymothoa exigua*) is a parasitic crustacean that, true to its name, first devours a fish's tongue before then latching on to the remaining stump and serving as a new permanent tongue for the remainder of the fish's life (see Figure 5.1). Brusca and Gilligan (1983), the biologists who first documented this creature, say it is "the first known case (in animals) of functional replacement of a host structure by a parasite." In this case, we have something further along the symbiosis continuum between parasitism and mutualism. That is to say, *after* the tongue is removed, the fish and the parasite depend upon one another mutually, although had the parasite not destroyed the fish's natural tongue the fish would never have needed the parasite. (In this sense, the parasite is a bit like a glass-repair company that throws rocks through your windows with their business card attached.) At any rate, in this case as well as in more mutualistic symbioses, there are items that simultaneously serve widely differing functions for different individuals—the louse's body serves itself in many regards, and it serves the fish as a tongue. Similarly, although liver flukes and oncogenes don't serve their hosts, they still serve themselves in many ways. And while clay crystals and the *in vitro* antibodies of the immune system don't even have hosts to serve, they also may be seen as serving themselves. In order to know for sure, we'll need to have a theory of what counts as the kind of self that might be served.

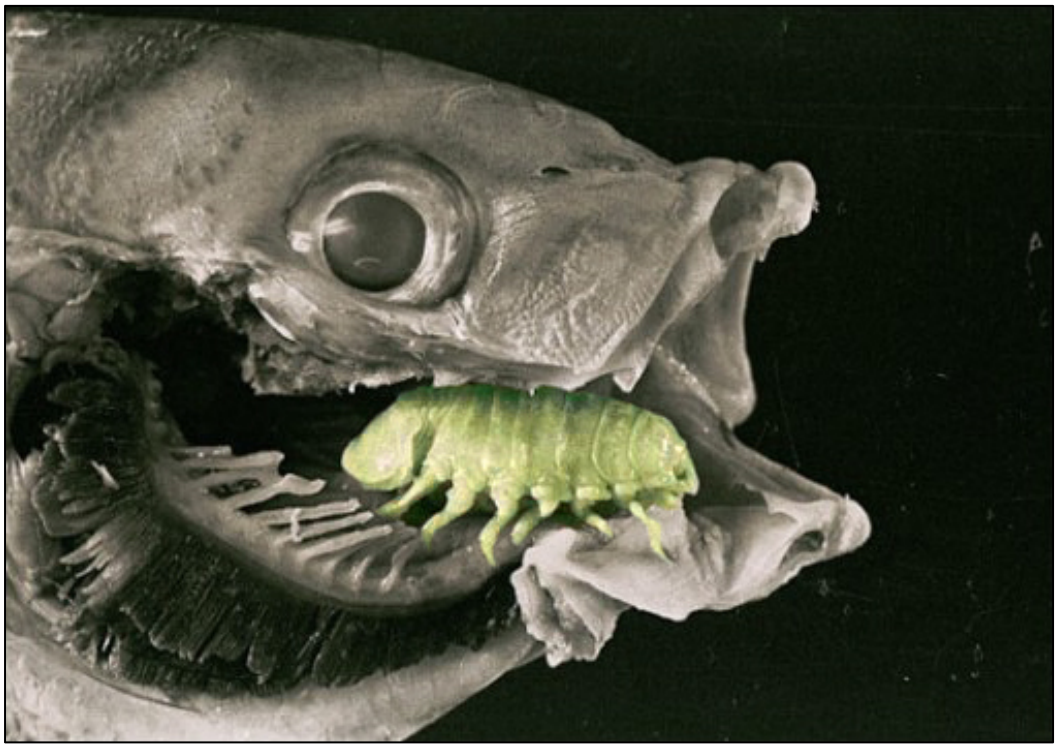


Figure 5.1: *Cymothoa exigua*, the tongue-eating louse (center, colored) serving as a tongue in the mouth of a fish. (Image credit: Matthew Gilligan, Savannah State University).

Segments of DNA in an organism's genome are referred to as "selfish" or even "ultraselfish" if the effects they have are, as far as we can tell, either invisible or even mildly damaging from the perspective of the organism, yet nonetheless advancing the cause of their own proliferation. Mere hitchhikers that sit on the genome and are passively copied once per generation—sometimes called "junk" or noncoding DNA[216]—are usually not cited as being selfish, since they are transcriptionally innocuous and only mildly proliferative, but they do share a similarity with what is normally called selfish DNA in that they are nonfunctional for the organism that bears

[216] However, this terminology is undergoing refinement as research progresses. More and more examples of segments previously believed to be noncoding have been determined to be functional, even despite not coding for proteins. At the very least, telomeres, centromeres, and segments that signal origins of replication all perform functions that help control the processing of genetic material within a cell. Still, there are plenty of segments that either do not appear to ever be transcribed to RNA, or that are transcribed to RNA that then goes on (so far as we currently know) to do nothing.

them and they are propagated, at least, across generations. Other more selfish segments have effects that further increase their own likelihood; transposons, for example, make multiple copies of themselves, even on the same strand—altering the genome typically with no beneficial effect for the organism's phenotype (which, through interaction with the world, determines the reproductive fitness of that genome). According to some calculations, through accumulation over the ages, transposons have come to make up as much as 50% of some organisms' genomes while still having no recognized phenotypic effect (see, *e.g.*, San Miguel *et al.* 1996).

Segregation distorter genes make up a slightly different class of selfishly behaving DNA. These are genes that actively and preferentially distort the ratio of alleles present in a population of gametes containing themselves. Here's the high-level description: When gametes (sperms or eggs) are created in an organism, the division process, called meiosis, normally results in four genetically distinct cells being produced from one of the organism's genetically "standard"[217] germ-line cells. One part of the meiosis process, called crossover, allows genes to be mixed and matched, introducing a fairly random element to the makeup of the four genetically distinct resultant gametes. Ideally, after this process has happened many, many times, the populations of resultant gametes that carry the various alleles that exist at any particular locus on the genome will be of roughly equivalent size. If there were two possible alleles at a particular position, then, after meiosis, half the gametes typically would carry one of the alleles at that location, and half would carry the other. Segregation distorter genes interfere with this ratio by "murdering" or otherwise disabling the gametes that have their competing allele, prior to the gamete-coupling process of sexual reproduction. In one well-studied version, these genes commit their attempted microscopic genocide by creating a kind of toxin to which they themselves have the antidote. As many of the gametes with competing alleles

[217] It is interesting to note that this view of "standard" is being challenged now, in the age of mass genomics, as there appears to be a considerable amount of genomic variation between the cells in the individual bodies of large multicellular organisms.

die off, the segregation distorter genes' own numbers see a relative increase, which expresses itself as a higher likelihood that offspring of the organism will carry the segregation distorter genes, rather than the newly dispatched alternative allele. In another version, the segregation distorters induce adverse mutations in the swimming apparatus of the sperms that carry their competitors, causing them simply to be less likely to reach an egg cell, with the same ultimate result (Burt and Trivers 2006).

As with the heritable variations in clay crystals and the selected antibodies of an *in vitro* immune system, these selfish-genetic phenomena are often cited as counterexamples to the SE analysis, since the case consists of selected effects that most authors seem unwilling to call functional (*e.g.* Manning 1997; Boorse 2002; Lewens 2004). Manning says "This seems to be a paradigmatic case of selection; having the trait of being a segregation distorter increases the chances of a bit of genetic material's being passed on through generations as compared with other genes without the trait . . . None the less, biologists do not typically regard [segregation distorters] as having the *function* of disrupting meiosis" (Manning 1997, as also cited in Lewens 2004).

And again, as with the previous two examples, I don't find the intuition (that Manning cites as coming from biologists) to be particularly persuasive. The behaviors of segments of selfish DNA, such as transposons and segregation distorter genes, certainly are not functional for the organism that they reside in, but it doesn't seem unreasonable to consider them functional for *themselves*. To take that perspective, though, one may need to give up one's assumptions about what counts as a self. A particular toxin that kills gametes might be functional for the segregation distorter gene, or a particular cancerous behavior in a cell may be functional for the oncogenes that cause it (or perhaps for the rogue cells that contain those mutant genes), in spite of their impact upon the larger organism that hosts these mechanisms, if those entities can benefit. So the question is one of identity—functional for whom?—and that is a question that can only be answered with a theory of

identity and value. Until we understand those topics, all three of these purported counterexamples against the SE analysis should be deemed inconclusive, as each is based upon their authors' intuitive assumptions about the relation of functions to certain beneficiaries.

Doubles and Initials

I think there is a very good reason for the preceding difficulty in finding false-positive counterexamples to the SE analysis: the things that have been selected by natural selection—including entities that are standardly considered to be traits of organisms, as well as naturally-selected phenomena that are not usually so considered—may in fact all be functional. If they are, however, I don't take it to be the case that they are functional *because* they were selected, but rather vice versa—I would say that they were selected because, among other things, they are functional. Their functioning played a role in their selection by helping themselves or their bearers (and thus, by proxy, themselves) succeed in their environment (see also Bigelow and Pargetter 1987).

Some things that are functional, however, are not the results of selection (that is, there are false negatives given by the SE analysis). The main category of false negative counterexamples aimed at the SE analysis is what is commonly referred to in the functions literature as "doubles"—organisms that appear magically and instantaneously rather than being produced by natural selection but that have the same structure as their natural counterparts (though any kind of a first occurrence of an animal or trait—call these "initials"—would also, in principle, fit the mold)[218]. Boorse introduced the idea of a double with what has come to be referred to as "instant lions".

[218] These kinds of examples may seem to be only boundary cases, but they are not. While it is difficult to draw out the meaningful differences in a process of gradual evolutionary change, some examples of single-mutation beneficial effects can occur. For instance, a member of a lineage of domesticated capsicum might in one generation produce mutant seeds that grow into new plants with a different color of their fruit. Farmers—and the markets that demand their products—may appreciate the new color, thus immediately kicking off a process of artificial selection for the new trait.

> Suppose we discovered, for example, that at some point the lion species simply sprang into existence by an unparalleled saltation. One would not regard this discovery as invalidating all functional claims about lions; it would show that in at least one case an intricate functional organization was created by chance. (Boorse 1976:74)

In other words, if our intuition aligns with Boorse's, the parts of instant lions and other "hopeful monsters" (Goldschmidt 1940; Gould 1982)[219] are functional despite lacking an evolutionary history. Their legs are for running, their noses for smelling, and their teeth for gnawing; and that is what these parts are used for, just like those of ordinary lions.

Millikan notes this challenge to her SE analysis and responds, as she admits "rather brazenly", that "such cases are like the case of fool's gold" (1989; see also Millikan 1996). She considers instant lions to be things that, were they to exist, would convince us of their function-bearing nature, yet that would not be functional in the slightest.[220,221] There is no easy way to argue against an opinion like that, but she offers another argument:

Exaptations make for another class of examples: instead of a change in the organism, we might have a change in the environment. A spandrel, a nonaptation (that is: a trait that has not been selected for), or a trait that had come to be vestigial could, with a relatively sudden change in environment, quite quickly come to serve an entirely new function (Gould and Lewontin 1979; Gould and Vrba 1982). This is not uncommon and, in terms of their relationships with the new environment, such traits, when they first appear, should be considered "initials".

[219] See also the literature on another kind of double, the so-called "Swampman" who is fortuitously generated (by, say, cosmic coincidence or lightning strike or radiation burst or whatever) as an identical model of an individual person (who might be disintegrated nearby at just the same instant). The thought experiment has been used to argue that Swampman would lack intentionality—would have no historically grounded semantic connection to the world either through ontogenetic (personal) history (Davidson 1987) or phylogenetic (evolutionary) history (Millikan 1984; 1996); more or less the same argument applies equally to historical notions of function (Millikan 1984; 1996).

[220] In other words, she would have us take these apparent-functions to be illusions. But she offers no perspective-changing test by which we could recognize their illusory status.

[221] In a personal communication, Douglas Hofstadter has pointed out to me that this suggestion of Millkan's sounds quite analogous to, and equally problematic as, Searle's (1980) claim that if a machine or mechanism such as his famous "Chinese room" passed the Turing Test, it would still have no semantics and no understanding and no intentionality (see also Hofstadter 1980).

> For example, your randomly created double exhibits no purposive behaviors and has no purposive parts because there is no way that any of his/her states or parts could be defective or might fail. That creature of accident, wonderful as he or she may be, falls under no norms (Millikan 1989).

This strange argument sounds not only theoretically biased—that is, it appears to be based only upon the SE analysis' claims about what constitutes norms (*i.e.*, selection)—but it also appears to be false. I imagine that, were an instant lion to appear and behave in all respects as a normal lion does, we could easily diagnose it with liver failure, arteriosclerosis, or ventricular fibrillation as it aged, based firstly upon comparative norms that we derive from its striking similarity to normal lions (and other homologous mammals; see Amundson and Lauder 1994) and, secondly, upon evaluative norms that we derive from the contribution of its liver and heart to its survival up until its declining health, and our expectation that a repaired liver, vasculature, or heart—or at least a *functionally* replaced one, in the case of an artificial transplant—would contribute similarly. There are a number of clear norms available to work with; Millikan just chooses to ignore them when she claims the creature "falls under no norms" (see Neander, 1991, for a similar argument that also dismisses any norms except historical SE norms)[222].

[222] "Some theories which imply that instant lions (and piggyback traits) would have proper functions do not capture the distinction between what an item does and what it is supposed to do, and so they do not describe a notion of a "proper function" that is capable of generating these biological categories which embrace both interspecies and pathological diversity." (Neander 1991)

Sorting Processes

Imaginary doubles are sometimes considered unfair examples because they don't naturally occur in our world.[223] But if the comparative norms of historical performances do indeed bestow a function upon a current item, as Millikan claims, then another class of real counterexamples, based also upon norms of historical performance, may promote a more convincing concern.

Tim Lewens describes *sorting* processes where "there is variation across a collection of items, [as well as] differential propensities among the items to survive some kind of test, but no reproduction" (2004). This process, representing one version of what Godfrey Smith (2009) calls *marginal* natural-selection, is a close cousin of full-fledged natural selection but, in terms of comparison against the SE analysis, differs crucially in its lack of reproduction. One example of Lewens' sorting processes is the screening, done by drug research companies, of millions of randomly generated molecules, testing for some effect. Another is granular convection—the process by which Brazil nuts, for instance, end up in the tops of containers of mixed nuts after undergoing the shaking that attends their packaging and lengthy shipment. A third is the sorting of pebbles, by size, performed by the waves on a beachfront. Lewens suggests that the "successful" items that emerge from such a sorting process can be said to have the function of causing that effect. In other words, the feature of pebbles (being light in weight) that allows them to be pushed highest onto the beach seems to function in helping those lightweight pebbles get pushed highest up onto the beach. Lewens says, "If selection can give genuine functions to eyes, then sorting processes can give genuine functions to stones on the beach" since "both processes support the

[223] I happen to think that the notion of doubles makes a fair case. If, one day, we are able to invent or discover new organisms by inserting whole-cloth, engineered strands of DNA into the nucleus of an egg cell from an actual organism, much like today's cloning technologies, then we really will be faced with explaining the functions of traits that appeared by what Boorse called an "unparalleled saltation". Even randomly generated strands might work, if we try a vast number of variations in the hope that just one might be a viable, if strange, new organism.

three connotations widely thought to be the marks of genuine teleology . . . [they] explain the presence of the functionally characterized item, . . . express normative demands on the item, and . . . allow a distinction between function and 'accidents"' (Lewens 2004; see also Dawkins 1983).

The types of norms involved here are, to my eyes and Lewens', very similar to the kind of comparative norms that Millikan cites in favor of the SE analysis. That is, the norm is derived from having succeeded at something historically, allowing us to reason that the item may continue doing the same thing in the future.

Using only norms of historical performance leaves us with two problems for the SE analysis, then. First, it raises the problem of dealing with doubles—that is, historical norms do not count the traits of instant lions as being functional despite even Millikan's intuitions that those traits would convincingly *appear* functional—and, second, it raises the difficult question of whether or not the survivors in Lewens' sorting processes deserve to be granted SE functions the same way the survivors of selection processes supposedly do—that is, historical norms seem to count Brazil nuts as being functional for going to the tops of containers. I will have more to say about the normative aspect of the SE analysis shortly, but first I have a further and perhaps more serious worry to express about the constitution of doubles and their initials.

Shadows and Residues

The SE analysis assumes that a function is a property of an item that has a certain historical story that can be told about it—not only must the item have been copied (Millikan 1984) but, in some versions (*e.g.*, Millikan 1993) it must have been copied at the expense of something else not being copied. The most troubling concern I have with this theory is that the reference to "history" amounts to extraphysical or metaphysical claims not too different in kind from old-fashioned,

enchanted vitalism or backwards causation (though, as we'll see, in this case the problem is *disconnected* forwards causation). An SE analyst who embraces the story that an item is granted some feature due to the selective history that produced it, and yet who refuses to accept that a double—a physically identical structure—has that same feature, is left to choose one of the following two unappealing metaphysical options.

First, they could assume some extraphysical substance or process or structure—something beyond what is measurable by our physical instruments—some magical *residue*, not made of the kinds of particles or energy we know the world to be made of, that is somehow left behind in an item with a selective history, but no trace of which is to be found on a *physically identical* item that does not have a selective history. For the sake of amusement, let's call such a residue a "functino" and imagine it to be a particle that carries the property of functionality. Option one is that the SE analyst needs to believe in functinos.[224]

Second, if they refuse the idea of an identifiable residue, then they could assume that functions are a causally irrelevant property—an epiphenomenon in the strictest sense, not just a shadow but an invisible shadow. On this option, the analyst would take the view that functional items and their doubles have a differing nature—that is, one has a function and the other does not—despite the fact that they are truly physically identical. The problem with this is that since they are physically identical, their causal futures (*ceteris paribus* with respect to the environments they are embedded within) are likewise going to be identical. The only way I see to reconcile identical causal behavior with "differing natures" (and without being contradictory with our modern understanding

[224] To extend the absurdity, it is worth noting that if one did believe in functinos, one would also have to specify some further properties of their behavior. In particular, one would have to clarify how these particles avoid being incorporated into the rocks and the clouds and so on, so that those objects don't magically spring to life. Is there a reservoir somewhere where the unused functinos reside? One would also have to clarify the special process by which functinos are imparted (presumably from ourselves?) into our designed artifacts. Would the very act of creating an artifact cause us to lose our own functinos, becoming ever more purposeless as we impart functions to these objects? Are assembly-line workers at the greatest risk of becoming hollow shells whose own parts eventually will stop functioning? Of course this is all ridiculous. There are no functinos, no purpose-bits in the world; purposiveness is a property not of particles but of relationships, of organization . . .

of causation in physics) would be to suggest that the difference is a non-causal one. Option two is that the SE analyst needs to believe that "having a function" is irrelevant—it *doesn't do anything*.

The troubling aspect of this dilemma (and I presume that, if pressed, the modern SE analyst would, like most of us, reject both of its horns based on materialist assumptions) is just one more reason that I am unwilling to believe in "proper functions". If we give up the notion that a function is a property that somehow inheres in an item, if we recognize the illusion of function constancy, and if we instead see the functioning of an item in terms of some relationship in which the item is held, rather than in terms of a historical, non-physical, supposedly property-granting process, then the problem of choosing between functinos and causally irrelevant functions simply disappears. Along with it go the problems of initials and doubles and vestiges, and also the malfunction fallacy, which is the next topic we're going to look at.

The Malfunction Fallacy

How could an item *fail to* function if it didn't *have* a function? That's the major premise of the malfunction argument, a cornerstone defense of the SE analysis (Griffiths 1993; Millikan 1989, 1993; Neander 1991, 1995). The idea, of course, is that a malfunction is simply the failure to meet the normative standard set by an actual function. A malfunctioning item just doesn't quite operate as it "should" or as it is "supposed to". If you recall, Millikan employed a version of the malfunction argument to defend her stance on doubles, claiming that the parts of doubles are simply *unable to be defective* because there is no historical norm to compare them to and so, according to her, since they can't malfunction, they cannot have functions.

It is sometimes said that no other theory is able to account for malfunctions and that this is a strong reason to prefer the SE analysis. Millikan puts it this way: a "fact about function categories is

that their members can always be defective—diseased, malformed, injured, broken, dysfunctional, *etc*.,—hence unable to perform the very functions by which they get their names. [. . .] The problem is, how did the atypical members of the category that cannot perform its defining function get into the same function category as the things that actually can perform the function?" (1989). According to her, the answer to this question is that the defective items got into the category by having the same selective histories—by being members of the same "reproductively established family"—as the functioning items and so, despite the defective items' current capacities or propensities (or lack thereof), they *have* functions.[225]

Now, I have claimed that even items that are functional *don't* have proper functions—that the idea of a proper function is itself an illusion—and it follows from that claim that nonfunctional items certainly wouldn't have them either. In addition, in the previous subsection I cast doubt on the idea that historical performance can provide the norms underlying functioning. But then, on such an account as I am giving, what sense can we make of Millikan's intuition that items might fail to fulfill their functions? How *else* might atypical, nonperforming members get into function categories?

Well . . . by analogy. A simple alternative (and one that Millikan acknowledges but discounts) is that dysfunctional items are described as such only because of their *resemblance* to items that we know to function in those ways. If functioning items are taken to have functions because they commonly perform those functions (that is, via the illusion of function constancy), and if they really are grouped together into functional categories (such as the category of hearts) because of various likenesses—gross morphological likenesses as well as functional likenesses[226]—and not because of any shared type of history, then it is no great leap to group other non-functioning items

[225] Neander (1991, 1995) and Griffiths (1993) make much the same point. Neander says, "Items that are dysfunctional are dysfunctional precisely because of their incapacity to perform their proper function" (1991).
[226] See Amundson and Lauder (1994).

along with them, items that share many of the same gross morphological features and which, we can reason, *would* also share functional capacities if only some small distinguishing (function-breaking) feature hadn't been so. (Here, we might recall the non-functional designed items on page 245.) What I am suggesting is this: Dysfunctional items are called "dysfunctional" only because (or only *when*) they remind us of functional items. Neither group actually has a function; there is no norm, universal to both and historically established, that one group attains and that the other falls short of. Instead, one group—the functional items—by the very functioning of its members, provides a norm against which the (similar-looking) dysfunctional items are measured and found to be deficient.[227] It is not that these dysfunctional items cannot perform a function that is properly "theirs"; it is only that they cannot perform a function that they look as if they ought to be able to—a function that we imagine them to have. The groups are subject to being compared in the first place simply because of various resemblances.

If this alternative were to be the case, then we would expect to see a spectrum in our function attributions that varies with level of resemblance: at one end, the more a functionless target item looked like a functioning version of an item (*e.g.* an actual heart), the more likely we would be to consider it dysfunctional. For instance, consider a heart that resembled other hearts in most respects including not only most details of its gross and fine morphology, but also its location in an organism and various types of connectivity with other organs. But suppose this heart had a hole in its ventricular septum, and suppose we widened that hole to the point at which the heart could no longer produce enough of a pressure differential to pump blood. At this point we would surely call this heart dysfunctional; it is a broken heart. At the other end of the spectrum, the less an item resembled a functioning version, the more likely we would be to consider the new item to be a member of a different kind all together. For instance, consider a calf, stillborn with a tumor

[227] Aristotle, in his *De Anima*, put it this way: "The eye is the matter of sight; if sight is lost, it is no longer an eye, except homonymously, in the way that a stone eye or painted eye is."

composed of what appear to be liver cells that grew in the space where a heart should have been[228]. Despite the mild similarity of its location in the calf, and despite the fact that it may have arisen from a stem cell that was originally produced on a path towards creating a heart, we would undoubtedly consider this item a tumor or a developmental oddity and not any kind of a heart[229]. Between these extremes we might find another stillborn calf with a solid, chamberless lump of striated cardiac muscle that has nothing like the form necessary to pressurize blood flow into the vasculature of a body.[230] This third item seems more like a heart than the liver-cell tumor, but more unlike a heart than the one with the perforated septum; we might find ourselves torn between calling it a dysfunctional heart and calling it a non-heart[231].

The question for the SE analysis is this: just where, in the developmental process from gamete to embryo to fetus to calf, might an alteration—a mutation, a developmental defect—be considered extreme enough for us to say that an item is a non-heart; and where might we consider it still merely to be a dysfunctional heart? Just how much similarity is required to consider an item to be a member of a reproductively established family, and what level of differentiation is required for an item to no longer be? The SE analysis has no way to give clear answers to these questions, for selective history has no bearing on developmental changes. While Millikan claims the SE analysis to be the best (or only) account of how to differentiate malfunctions from functions, firstly, this is false since other norms are indeed available and, secondly the SE analysis fails to differentiate malfunctions from some non-functions.

---

[228] This may be a developmental impossibility, since a calf fetus without the rudiments of a heart would likely be unable to produce much of any type of tissue, but the imagined scenario is instructive nonetheless.

[229] To exaggerate the case, we can imagine a stillborn calf fetus so badly formed that the entire thing is entirely unrecognizable; the SE theory seems to suggest, oddly, that some blob of cells in the mass of undifferentiated flesh still "has the function" of pumping blood simply because it has a historical relationship to its parents' hearts.

[230] These examples are fantastical, since a beating heart is a practical and functional requirement for keeping a fetus developing in the womb. However they are easily understood and, while changing the example to a wing or an eye would work as well, it wouldn't fit with the theme of hearts.

[231] McLaughlin notes that if a spontaneous Porsche (a double) occurs with a broken axle, "we would be hard pressed to say whether it is a malfunctioning car or just not a car at all" (2001:49). I can see how the case may not convince everyone; but variations of it may indeed be more convincing.

This is one way in which we run into trouble determining membership in a reproductively established family; we'll look at another way in the next section.

Indeterminacy

As the SE analysis has developed over the years, one of the central debates among its proponents has been about just *how much* natural selection is required for a function to be granted. We already indirectly faced this question when talking about doubles earlier: when exactly do SE analysts grant a function to the traits of a newly created instant lion lineage? As we've been told, it is certainly not immediately . . . but is it enough to say they have functions by the time the lion has created its own cubs? Or do the original instant lion's traits never have functions, but perhaps its cubs' or its grandcubs' traits do?

The reason nobody knows how to answer these questions is that "amount of selection" is a *graded* phenomenon, while the possession of a "proper function" is taken to be black-and-white. Suggestions for where to make the distinction have included the entire history of the trait, only the most recent history of the trait (Godfrey-Smith 1994; Millikan 1989b), or a combination of distant history along with present-day "continuing usefulness" of the trait (*e.g.* Schwartz 2002)[232]. Not only do these proposed answers disagree with one another, but each one itself also fails to draw a fine line, and so having a function, on any such account, is still either a matter of degree, or of some kind of unspecified subjective determination (either of which countermands the SE analyst's black-and-white notion of having a proper function).

[232] An alternative way to draw the line says that a history of selection, while it must have occurred to create a trait, does not confer a function; instead the trait's function owes its status to the trait's being currently causally disposed toward being replicated (Bigelow and Pargetter 1987; Griffiths 2009). This view is called the dispositional theory, and I will address it in detail in the next section.

A similar issue of indeterminacy has been critically raised by Allen (2002), who asks us to consider the existence of functions to be as blurry as the underlying concept of traits itself (see also Gould and Lewontin 1979; Hardcastle 2002; Kauffman 1971).

For one thing, traits are hard to bracket into a well-defined category across time. A heart and its ancestors are never entirely identical—near relatives, only a generation or two apart, may be *relatively* similar, but such is the nature of gradual change; ancestors and descendants separated by many more generations may differ more substantially, and at some point the differences will be so large that we would no longer consider those individuals at one end of the family tree to be "the same" as those at the other end. It is thus unclear what criteria Millikan and other SE theorists might have us use to define a "reproductively established family". Are the sometimes widely differing traits of a ring species[233] all "in the same family"? By what measure? If we were to use the members' function itself to establish traits (and thus the family), then the definition of function in terms of a reproductively established family would become circular.

For another thing, individual traits are hard to pin down even within a generation or an instant. As was discussed earlier (p. 277), it doesn't seem possible to find distinct joints along which to definitively carve an individual heart from its body. Prior (1985) points this out also, saying "Organisms do not come to us divided into parts and with labels on those parts. *We* (or more strictly anatomists and physiologists) analyse them into those parts and attach appropriate labels." Function plays a major role (perhaps second only to morphology) in that reasoning.

[233] A ring species is one that has a geographical range that circles or nearly circles the globe, yet in which there is variation all along that range such that each set of neighbors may interbreed but those at farther, possibly overlapping, ends from one another cannot.

Allen suggests that if we can offer no clear answer as to what a trait is or as to what constitutes being a member of a reproductively established family, then SE functions, conceived of as properties of traits that are features of those families, must inherit this imprecision.[234]

The Chicken and the Egg

Many authors have pointed out a paradoxical-seeming circularity in the application of the SE analysis. The point is a worthwhile one, but one that needn't be made anew by me, so I will simply borrow from other writers here. Griffiths makes the case best and I find it worth quoting him at length[235].

> Millikan has correctly pointed out that in order to study the biological functioning of an organism, biologists must identify where one organism ends and another begins, must distinguish the functioning of that organism from irrelevant causal processes in which the organism is caught up, and must identify and exclude pathological features of the organism. She suggests that biologists determine whether something is part of an organism's biological functioning, and thus solve these problems, by determining that it has a selected function (or has been exapted to support a selected function). But this suggestion generates a paradox, because the first step in determining whether something has a selected function is to analyse the contribution it made to biological functioning in the past. To show that oddly-shaped sperm have the selected function of interfering with the sperm of rival males, it is necessary to show

[234] However, this problem can be elided if things simply don't have functions.
[235] It is worth pointing out here that this Griffiths (2009) who is arguing against the selected effects theory is the same Griffiths who sixteen years earlier (1993) had offered a selected effects theory. Something must have changed his mind in the meantime.

> that these sperm increased the fitness of ancestral males that produced them by interfering with the sperm of rival males. But either we can establish this without knowing the selected function of the sperm of those ancestral males, in which case we could do the same for living males, or we have to know their selected function in those ancestors, which means looking at still earlier ancestors to discover this and so on *ad infinitum*. (Griffiths 2009, p. 18)

The regressive inquiry that Griffiths is concerned with is problematic not only because of its regressive nature[236], but also because biological creatures undergo constant gradual change, and so it wouldn't take too many layers of chasing reasons back in time before one found oneself simply looking at a different organism than where one started, and thus one would simply be unable to answer questions about the original organism of inquiry.

A focus upon the role of natural selection in one or another analysis of functions is to be expected. There is a clear and strong correlation between things in the biological world that have functions and things that have been naturally selected. And it is convincing to think that the bulk of, if not all of, the purposeful items in the world are the results of natural selection. But this is only to say that natural selection makes teleological things, not that it makes things teleological.

It is clear to anyone familiar with evolutionary theory that the reason a biological item is there—why it exists—has to do, in part, with the fact that its ancestral versions functioned properly. As Griffiths points out, this is because function logically precedes selection—an item can be selected

[236] Prior (1985:321) makes a similar point: Evolutionary biologists depend upon the delineation of traits performed by anatomists and physiologists in order to explain relative fitness (see also Allen 2002; Amundson and Lauder 1994). In Griffiths' terms, she is pointing out that we needn't follow the evolutionary regression *ad infinitum*; we can establish functioning from present facts.

for only if it functions (see also Amundson and Lauder 1994; Bigelow and Pargetter 1987; Cummins 2002; Davies 2001[237]). Later, Griffiths summarizes.

> "If biologists needed to know why something evolved before they could describe its form and function then they could never get off the ground." (ibid, p. 26)

Now, we must admit it is not entirely circular to say both that:

> C1) An item's functions are the effects for which it has been selected, and

> C2) Selection for an item occurs because of the effects of its functions.

Eggs beget chickens, which then beget eggs . . . But a cyclical analysis of this sort must ground out somewhere. There must be a base case. Which came first?

Functioning is Normative

The SE and CR analyses take diametrically opposite positions with regard to the normativity of functions. While the CR analysis imagines function to be a completely non-normative notion, and its proponents commonly ridicule the idea that nature could be in any way normative or subjective, the SE analysis prides itself on its claim that it gives the only possible analysis that can account for normativity in nature.

[237] "We cannot discover the selected function of any trait without first knowing its systemic function. If we do not know the systemic function of a trait, we have no guide with which to seek historical evidence for the claim that this trait was selected for the specified functional task." (Davies 2001, p. 55)

It should be no surprise that I side more with the SE viewpoint in this matter, since I've already argued that the CR analysis is a closet-teleological view that superficially appears value-free in its generic form, yet that surreptitiously smuggles in value-laden teleological terms during individual analyses. However, as it turns out, I agree with the SE analysis in spirit only. I appreciate the fact that it attempts to admit normativity, but it picks out what I think are entirely the wrong norms. In particular, the SE analysis puts function in terms of a comparative, not an evaluative norm; I find it to be *inappropriately* normative. History can play no role in function without inviting implausible metaphysical oddities, such as residual functinos or epiphenomenal functions.

Now, SE analysts seem satisfied with their comparative norms. Millikan (2002) points out that her term "proper function" was never meant to be an evaluative term. And, as was discussed above, it is comparative norms that buoy her view of malfunction, which she defines as not functioning as history would dictate. The way I see it, though, using comparative, historical norms leads us into a number of problems. First, they conscript the results of Lewens' sorting processes into the ranks of function-bearing items while excluding instant lions. Second, they provide us with a way to make sense of the distinction between functions and malfunctions, but only at the cost of not being able to distinguish between malfunctions and certain non-functions—a difficulty that also highlights the failure of the SE analysis to provide a non-problematic way to define its own central term, "reproductively established family". Third, and most importantly, they present us with the dilemma of either accepting the existence of causally relevant non-physical residues created by a history of selection or accepting the idea that functions might themselves be causally irrelevant.

All in all, I may sound negative about the SE analysis, and in particular about its normative claims, but I nonetheless find several insights from the SE analysis to be worth retaining. First, it is difficult to find convincing counterexamples: things that have been naturally selected do seem to be correlated with things that are functional, and if we are to offer a view that natural selection is not

the cause of that functionality, then this correlation is a fact that deserves some other careful explanation. Second, as I've emphasized, I agree with the centrally normative aspect of the SE analysis; I only disagree on the particular source of normativity invoked by it. Third, I think Wright's basic insight—that functioning occurs in items that, in some way, cause themselves—is crucial, though the devil is in the details, and we'll have to explore those details in later chapters. Lastly, I find Millikan's distinction between direct and derived function(ing) to have some resemblance to the view I will advocate. The relation she describes—in which a derived functional item is produced by a direct member of a reproductively established family—is not quite the way I'd like to describe it, but it is close.

C. Replication Dispositions

The relevant properties are seen either as those which contribute to an organism's current needs, purposes and goals . . . or those which have evolutionary significance to the organism's survival and reproduction.

—Ron Amundson and George Lauder (1994)

[The] decoupling of function from the processes responsible for its origination means that the specific mechanisms involved don't matter. Only the consequence matters, irrespective of how it was achieved . . . [This decoupling] helped [Darwin] to recognize that variations of structure and function that arise by accident can nevertheless be functional.

—Terrence Deacon (2013, p. 423)

The last of our three popular theories is the Replication Dispositions (RD) theory of function (also sometimes called the Propensity theory). The RD analysis is similar to the SE analysis in that both consider the capacities of traits that aid in the survival and reproduction of the trait-bearer to be the functions of those traits, but the two views differ in that the SE analysis confers functions upon traits whose ancestral versions had effectively contributed to their own modern existence, while the RD analysis confers functions upon traits, now, that will effectively contribute to the creation of their own successors. So for RD theorists, a function is *a disposition to contribute to replication*. The most commonly cited proponents of this view put it this way.

> The [SE] theory describes a character now as serving a function, when it did confer propensities that improved the chances of survival. We suggest that it is appropriate,

> in such a case, to say that the character *has been serving that function all along*. Even before it had contributed (in an appropriate way) to survival, it had conferred a survival-enhancing propensity on the creature. And to confer such a propensity, we suggest, is what constitutes a function. Something has a (biological) function just when it confers a survival-enhancing propensity on a creature that possesses it. (Bigelow and Pargetter 1987, emphasis added)

The RD analysis claims that the function of the heart is pumping blood because pumping blood is what the heart does now that will help ensure that the organism whose heart it is will be able to reproduce and create another organism (with a similar heart). It also differs from the SE analysis, because it is taken to be a token-, rather than a type-, theory—that is, it makes claims about *this heart*, rather than *the heart* as a category, a fact that allows the RD analysis to refrain from granting functional status to individual malformed hearts that do not confer a survival-enhancing or reproductive propensity (despite their possibly being members of a reproductively established family). The RD analysis simply does not fall for the malfunction fallacy. Another consequence of the forward-looking nature of the RD analysis is that it deviates from the SE analysis in cases of initials and doubles, such as Boorse's instant lions. In these cases, the RD analysis claims that the hearts and other parts of these animals do function, as long as the animals are reproductively viable.

Applied to our other two base cases, the RD analysis claims that the cog and the stone paperweight, and in fact most artifacts, have no functions. Cogs, and the machines they are part of, typically have no dispositions toward being replicated. Similarly, a stone paperweight leads a lonely existence and is unlikely to produce progeny of any kind. While copying from a blueprint or a prototype does play a role in generating some mass-produced artifacts, it seems that this type of procedure should not count as replication (in the RD sense) since the functional items in question,

while they are products of such copying, are not themselves likely to serve as prototypes for further copying. That is, they themselves are not being replicated; they are just siblings—the *results* of replication of a single prototype parent. Some artifacts do go through a process of technological evolution in which later versions are copies based upon, but possibly improved from, earlier versions (Darden and Cain 1989; Griffiths, 1993). However, for one thing, this is not necessary (many artifacts don't play such a role, such as our stone paperweight or Cummins' (2002) Rube Goldberg-style gate-opening device[238]—they are limited-edition designs that serve possibly unique functions). For another thing, the copies that do come about are generally functional whether or not they have any tendency to be further copied.[239]

A Historical Note on Variants

A few variants of the RD view exist, but they differ from one another only marginally. Central offerings include Bigelow and Pargetter (1987) and Griffiths (2009)[240] but the survival-and-reproduction theory of function (call it "SR") is a close relative, too. Canfield (1964) and Ruse (1971) each offered an SR theory in which a biological trait has a function when it is an adaptation that contributes to the reproductive fitness of its bearer. Here, for example, is Canfield's version.

---

[238] Cummins (2002) describes a device he has constructed on his farm, from a wind-up alarm clock, a piece of string, and a latch which, all together, serve to open the gate in a field at a particular time. While our world has many, many more mass-produced devices, one-off designs such as this are not uncommon either.

[239] Though the bulk of their paper is about biological functions, Bigelow and Pargetter (1987) claim to have accounted for artifacts as well. However, I am sorely unconvinced by the paragraph in which they do so: They alter their view, momentarily, repositioning the theory in terms of a "propensity for selection", instead of a propensity for survival (as they write throughout the rest of the paper) and then suggest that, in the case of artifacts, selection occurs at the moment an item is made from a blueprint or prototype, not later. This claim departs widely from their account of biological functions though since, items made from blueprints or prototypes usually do not have a propensity for survival or replication at all. And, as noted in the main text above, some one-off artifacts are neither made from a prototype or blueprint nor serve as one. Additionally, Bigelow and Pargetter don't seem committed to this altered notion, since, as soon as they conclude that paragraph, they cease to ever make mention of a "propensity for selection" or even just selection; they revert immediately to discussion of propensities for survival. Because of the discrepancy between their biological theory and their artifact theory, I will treat their extensively laid out biological theory as authoritative of their view and largely ignore the artifact paragraph.

[240] See also Mills and Beatty (1979) for a "propensity account" of fitness that underlies these kinds of considerations.

> A function of *I* (in *S*) is to do *C means I* does *C*; and if, *ceteris paribus*, *C* were not done in an *S*, then the probability of that *S* surviving or having descendants would be smaller than the probability of an *S* in which *C* is done surviving or having descendants. (Canfield 1964, p. 292)

Griffiths makes a similar SR-like claim when he works to differentiate his RD view from SE analyses[241].

> Rather than focusing on causal capacities that featured in past episodes of selection, we should focus on causal capacities that contribute to survival and reproduction (survival value). (Griffiths 2009, p. 24)

While, give or take a detail or two, the SR analysis in some ways approximates the Goal Contribution account and the Valuable Effects account that we'll see shortly—and, in some other ways, it appears very similar to Wright's formula—I nonetheless see it as most closely related to Bigelow and Pargetter's RD analysis, since both are strictly biological theories related to replication and, in each, it is the contribution towards creating future progeny, determined by present causal capacities, that counts as a biological function.

[241] To clarify: Griffiths (1993) offered an SE view of function, closely aligned with Millikan's, but did an about-face with his (2009) RD analysis.

Counterexamples

As with the previous two theories, an analysis of some counterexamples can help to shed light on the rough patches. The central class of counterexamples that bears meaningfully on the RD (and SR) analysis is those situations that involve sterility.

We would like to say that the heart of a mule has the function of pumping blood, even though such hearts can have no disposition to replicate. The example affects not only categorically sterile animals, such as mules, but also creatures with acquired sterility that once had the disposition to replicate but then lost it.[242] In these cases, our intuition (that, for instance, such an organism's heart is still functional) aligns with Cummins' CR analysis (since the heart plays a causal role in the mule's circulatory system) and with the SE analysis (since we might take the heart to be a member or a product of a member of a reproductively established family[243]) but it clashes with the results of an RD analysis.

Self-Reproduction

One attempt to account for sterile creatures, without resorting to an SE or a CR analysis, would be to amend the RD analysis and put it in terms only of survival, without reproduction. That is, we could take Bigelow and Pargetter's phrase more literally than they intended it and say that an item "has a (biological) function just when it confers a *survival-enhancing* propensity on a creature that

[242] Hardcastle (2002) also gives the example of the eyes (or any traits) of a sterile worker bee that will never reproduce. This example, however, suffers from a prejudice of drawing organismal boundaries at physical boundaries. Social insects whose reproductive delegates constitute a minority of the population are better seen as distributed organisms—a single animal with many partial bodies—rather than as sterile individuals (see also Dawkins 1982; Wilson 1971, 2009).

[243] There is room for interpretation here, since having both a horse heart and a donkey heart as ancestors doesn't necessarily make a mule heart a "copy" as it will likely differ from both its parents in some ways; but a less extreme version of this same concern affects the SE theory more generally since any sexually reproduced organism's organs will differ somewhat from its parents'.

possesses it" (1987, emphasis added)[244]. A number of authors in recent years have made just such a case, suggesting that a function is whatever contributes to "self-reproduction"—the ability of an organism (or a trait of one) to protect, preserve, repair, and rebuild the organism over its lifetime (see Christensen 1996; Christensen and Bickhard 2002; DeLancey 2006; McLaughlin 2001[245]; and Schlosser 1998). The ability to self-reproduce—also referred to by Maturana and Varela (1973) as *autopoiesis,* a notion I'll review in more detail in Part II—shares the intuitively purposeful-seeming, active, vitalistic nature that we see also in reproduction. On its own, though, a self-reproduction view is not clearly superior to the original RD analysis. It suffers from the same inability to classify artifact functions (aside from the occasional robot that plugs itself in to recharge its batteries, no contemporary artifact self-reproduces in any way) and it has its own counterexamples to contend with as well. For instance, most viruses are not self-rebuilding in the same way that cellular life is, yet they are replicators, and it would be strange not to consider their parts and the mechanics of their behaviors to be functional towards their reproductive ends.[246] Griffiths describes a couple more centrally biological cases.

> For example, the genetic and developmental mechanisms that underpin the failure of the mouthparts of mayflies to develop fully after metamorphosis to the adult reproductive stage make no sense when analysed for their contribution to the individual's ability to maintain its form ('self reproduction'). They make perfect sense

[244] Doing so would not work for Bigelow and Pargetter themselves if we were to take their artifact-directed revision seriously, since that revision requires a propensity for selection, which itself would require reproduction. They have a choice, I suppose: either keep the artifact-revision and retain difficulties explaining sterile animals, or abandon artifacts and possibly follow McLaughlin to explain functions in sterile animals.

[245] Interestingly, McLaughlin sees more of an affinity between his view and the SE analysis than between his and the RD analysis (he sees self-reproduction as simply replacing reproduction, leaving the basic form of Wright's or Millikan's schema otherwise little changed). On the other hand, because of his claims that self-reproduction is the basis of an intrinsic good (or value), I prefer to classify and further discuss his view under the Valuable Effects analysis that comes later in this chapter.

[246] We will review the status of viruses and their behaviors again later.

> as a contribution to reproduction. In several small Australian Dasyurid species such as *Antechnius Stuartii* a frenzied mating season is followed by a short period during which the male's sexual organs regress and their immune system collapses. Then all the males in the population die. The mechanisms that underpin this 'big bang mating' behavior (Diamond, 1982) obviously do not contribute to the capacity of individual males to maintain their form. But they do contribute to the life-history strategy by which these males maximize their contributions to future generations. (Griffiths 2009, p. 21)

And Peter McLaughlin points out another example.

> Parthenogenetically reproducing cecidomyian gall midges don't hatch their eggs outside their bodies but rather inside and are then devoured from within by their own growing daughters. But such lethal traits are said to have a biological function. (McLaughlin 2001, p. 79)

As these examples show, some functional behaviors in biology are detrimental to autopoietic self-reproduction yet beneficial to reproduction.

Despite counterexamples of this sort, I agree with the self-reproduction authors that their concept is in fact intuitively related to being functional. While neither the standard RD analysis nor the autopoietic RD analysis, on their own, seem to describe the full set of function attributions for which I am pursuing an explanation, together they suggest a notion that might help us get closer: survival *or* reproduction. I'll flesh out a version of this notion later, somewhat following Schlosser (1998), Christensen and Bickhard (2002), and DeLancey (2006).

On the Irrelevance of Natural Selection

I am not convinced, as Griffiths is, that only an *evolutionary* forward-looking account can explain function because "biological functioning must be understood in terms of reproduction, not only self-reproduction" (2009). I agree that biological functioning must (in some way) be understood in terms of reproduction, but, firstly, not necessarily over and above self-reproduction and, secondly, not under the conflated view that reproduction is, by default, evolutionary. It seems to me that evolution—conceived of as a set of *changes* in species due to *differential* reproduction in a population with heritable variation—is in fact irrelevant to functioning because functioning (or even "having a function") occurs in the present (see also Cummins 1975:745). The claims in Griffiths' paper seem to equate evolution directly with reproduction alone (despite the fact that Griffiths is well aware of what paradigmatic evolution by natural selection consists in[247]).

Of course *if* a function were a *disposition* for an item to survive and reproduce, as has been clearly stated by Griffiths as well as Canfield, Bigelow and Pargetter, and Ruse, then neither would the item need to succeed at replicating nor would it need to succeed at the expense of any competitors, nor would it need to change or have changed in any way, in an evolutionary sense. All it would need to do is be disposed toward not falling apart before it had a chance to at least set off a chain of events that will construct a copy of itself. Variation and heritability might stay entirely out of the picture, as would competitors. As I argued against the SE analysis, what possible physical, causal role could a competitor (past or present) play in the nature of an item? They might play a role in eliminating—outcompeting—an item, but they don't cause a successful item either to exist or to have the functional capacities it has.

[247] Perhaps this was a tactical error on Griffiths' part, instead of a logical or lethargic error, in that he may have been aiming his criticisms only at McLaughlin-style analyses. But it is difficult to read him that way, since the claim he makes is that his target is all "non-evolutionary accounts of biological functioning".

The survival and reproduction of individuals is certainly one prerequisite for evolution by natural selection. However, in the absence of other features, reproduction alone doesn't produce change or adaptation and doesn't engender either selection or evolution. Both can also logically exist without natural selection (Godfrey Smith 2009). At its core, the RD analysis—based only in replication—can be agnostic of change.

Replication and Self-Replication are Normative

When Canfield developed his SR view, described a few pages earlier, he first put it in terms of "usefulness" and then attempted to translate his way out of that commitment in order to dispense with what he saw as an unsustainably normative notion. He wanted a non-normative and thus non-teleological version of the same characterization—a characterization of the processes or mechanics underlying situations that we deem functional in which those processes could be seen as not being *for* anything.[248] According to Canfield, the translation of "usefulness" into the terms "survival and reproduction" successfully accomplished that job. I think his translation served only to obscure, not eradicate, the teleological norms implicit in his characterization[249].

Like Canfield, other RD theorists also appear to think that the use of such terms allows them to ground their theories in mechanical processes that are normatively innocuous (meaning, as per the discussion in Chapter I, that the concepts may be normative but—so the story goes—only in a comparative, not in an evaluative sense). The use of these concepts (replication, self-replication) is

[248] For the interested reader, here is Canfield's preliminary version, with an example: "A function of I (in S) is to do C means I does C and that C is done is useful to S. For example, (in vertebrates), a function of the liver is to secrete bile, and that bile is secreted in vertebrates is useful to them". (Canfield 1964, p. 290)

[249] This issue of translatability is taken up again in Chapter VII, following Beckner (1969) who argues, "teleological statements are not translatable into non-teleological ones."

precisely where I think the RD view implicitly imports an evaluative type of normativity into its analysis, while successfully hiding it behind the veneer of another, merely comparative normativity.

That is to say, terms such as "reproduction" or "self-reproduction" are easily conceived of as bringing with them comparative norms (as we saw with the SE analysis). There is an obvious sense in which reproduction and self-reproduction can succeed or fail in the same way as, say, a moon can succeed or fail at orbiting a planet; these processes can be compared to other instances (either historical or parallel) and either they just happen or they don't. If viewed externally by an analyst who is biased by the belief that value is not an inherent part of our universe, then nothing much appears to ride on the success or failure of that mechanistic operation. Any normativity that might be observed can be passed off as having arisen from the observer's comparative norm derived from other observations.

Attending only to that comparative pattern can, however, camouflage an underlying simultaneous evaluative normativity—that is, the fact that success or failure of reproduction or self-reproduction is, in actual fact, very important to the agent or entity that strives to reproduce or self-reproduce. There is an important sense in which success or failure is very much either good or bad for that entity. At the moment, this may seem to be an arguable intuition; however, one of the keystones of the theory advocated in Part II will be a foundation for this type of evaluative normativity.

Functioning Contributes to Dispositions to (Self-) Replicate

Of the three popular theories, the RD analysis is closest in form to the view I will take. I agree very much with the forward-looking perspective it takes, including the fact that it nicely captures our intuition that, as Bigelow and Pargetter put it, a thing has been serving that function all

along, and I find value in the observation that both reproduction and self-reproduction are behaviors that appear vitalistic.

The concerns that I raised were, for one thing, that the RD analysis is incomplete in that it borrows an evaluative notion without being explanatory of that notion and, for another thing, that individual versions of it (either reproduction versions or self-reproduction versions) have trouble making sense of the various counterexamples, such as the functions of artifact parts and sterile animals' organs or those of traits that are antithetical to self-reproduction. Millikan's insight—that an item may not cause its own direct copying, but might instead cause the copying of the thing that produced it—can offer help here. We might improve upon the RD analysis by suggesting that an item is functional whenever it is disposed to contribute to autopoietic maintenance of an item or to either replication of itself or replication of something else—*loosely*, its "user", in the case of artifacts, or its "owner", in the case of traits. If the stone paperweight or the machine that contains the cog contributes in some way to its user's disposition to replicate, then they fall under this extended type of RD analysis, while the machine or stone is no longer required to be a replicator itself. The amended analysis might seem like a two-part (disjunctive) theory, but I think a certain form of it can be seen as elegantly unifying.

D. Hybrids

Philosophical discussions of function have tended to pit different analyses and different intuitions against one another without noting the pluralism inherent in biological practice.

—Philip Kitcher (1993)

Hybrid theories of function that are based in some of the popular theories can be classified into three subtypes. Here, I will follow Davies (2001), who calls these three types of theories *unification*, *pluralism*, and *instantiation*. While none of these hybrid schemes fares any better than their individual components do, it is important to discuss them because defending one or another of them—most often the pluralist version—is the customary approach taken by most philosophers of biology these days. Generally, these philosophers find themselves persuaded that the SE theory accounts best for biological function, but they are also willing to allow that it doesn't account as clearly for artifacts, or for all the ways we look at biological functioning (see, *e.g.*, Amundson and Lauder 1994). In order to patch up this perceived gap, they adopt an additional provision that the remainder of our functional endorsements might, in one way or another, be attributable to perception of a structure such as Cummins' CR functions.

In what comes below I will describe the hybrid analyses, but I will mostly ignore counterexamples and specific criticisms which I find unnecessary, because these analyses are all plagued by the more general concern that the constituents from which they are constructed—the CR and SE analyses—are themselves in need of substantial revision.

Unification

Unification (or synthesis) is the idea that a suitably generic concept could subsume both the CR and SE analyses. If each were a subclass of a single superstructure of some sort, then perhaps that superstructure alone could account for function, while the particulars that differentiate the two subclasses could account for some of the nuanced differences in kinds of functions. Kitcher (1993), for instance, takes *generalized design*, which he portrays as incorporating both human and natural design, to be the central idea underpinning a generic concept of function (see also Dennett 1990). This notion subsumes the other two firstly because the SE account is based directly on natural selection as a source of design, and secondly because Kitcher argues that "when we attribute functions to entities that make a causal contribution to complex processes"—that is, when we attribute a CR function—"there is . . . always a source of design in the background" (1993: 390). As I stated earlier, in the discussion devoted to the design fallacy, the correlation between design and function is strong (and thus requires explanation), but it can be better accounted for by the fact that successful functioning guides design, rather than by the act of design granting proper functions. Kitcher appears to be not only smitten with the intuitions that underlie the design fallacy but also, like many other modern function theorists, captivated by the illusion of function constancy.

Walsh (1996) and Buller (1998, 1999, 2001) attempt to give a different unification that they refer to as "the relational theory". Under this analysis, a function gets its status from its contribution to fitness under a time-independent "selective regime". This formulation is meant to triply unify the SE, RD, and CR analyses, in that a function is a contribution to either past fitness (SE) or current fitness (RD and CR). However, it is a biologically focused analysis that (i) doesn't apply to artifacts; (ii) is still vulnerable to counterexamples such as Lewens' sorting processes, vestiges, doubles, and initials; and (iii) fails to explain not only what kind of a residue or structure past selection might

physically confer upon an item, but also what kind of structure or metaphysical residue future selection might be able to backward-causally confer upon an item in order to ensure that it is presently functional.

Pluralism

Pluralism, the most popular form of hybrid analysis, generally aims to bring the CR and SE analyses together by exploiting their differences rather than their similarities. The pluralist analyst believes that the SE theory describes one kind of function while the CR theory describes another kind of function, and that the two may overlap many times in biological cases, obscuring the fact that both concepts apply simultaneously. On this view, a cog need not have been selected for its effects to have its function, as long as it plays a causal role, and a seed (that never grows) need not in fact play a causal role to have its function, as long as it is the result of selection (of its ancestors) effects. But the function of a normal heart may both play a causal role and be a result of selection.

The proponents of such a model are many (Allen and Bekoff 1995a, 1995b; Amundson and Lauder 1994; Brandon 2011; Godfrey-Smith 1993, 1994; Hinde 1975; Melander 1997; Millikan 1989b, 2002; Preston 1998), though there are slight differences in all their analyses.

Perhaps the most central issue behind the pluralist view is that the SE and CR analyses are working towards different explanatory goals. Godfrey-Smith makes this clear:

> If it is claimed, for instance, that the function of the myelin sheaths round some brain cells is to make possible efficient long distance conduction of signals, it may not be obvious which explanatory project is involved—that of explaining why the

> sheath is there, or that of explaining how the brain manages to perform certain tasks. (Godfrey-Smith 1994)

This is a good point about the explanatory projects of biology, but it still does not require a plural view of functions; it just needs an explanation of both of these things, of which, one, both, or neither might involve function. In particular, I find it likely that function plays some role in both "why-is-it-there" and "how-does-it-work" types of explanatory project, but that it is not constitutive of a full explanation for either.

As I stated earlier, pluralism is unconvincing to me because both of the underlying analyses from which it is constructed are themselves problematic. For one thing, both analyses tend to take functions as things that are *had*, not things that are *served*, and conjoining them does not resolve this problem. For another thing, the SE analysis is subject to the malfunction fallacy, the design fallacy, and the metaphysical-residue concern, and the CR analysis is subject to its teleological smuggling concern, none of which are solved by borrowing either analysis into a plural view. And, for a third thing, borrowing both the CR and the SE analyses into a plural view brings along the unconvincing overbreadth that afflicts each individually (cf. their respective counterexamples in the previous sections of this chapter). Pluralism, in all its various treatments (see citations above), is only a loosely specified conjunction of its constituents that seems to properly assign its *inclusions* (such as the heart and the cog) but that seldom has attention paid to its *exclusions*. For instance, how does pluralism deal with Griffiths' piece of dirt stuck in a pipe, or with Millikan's rain clouds or with oncogenes (Boorse 2002)? The selected-effects analysis correctly says they are not functions, but in pluralism, candidates for functionhood must be rejected by *both* components of the theory in order to be fully rejected, and as we've already seen, the causal-role analysis has trouble rejecting these

kinds of examples.[250] I am not opposed to a plural view in principle, but firstly, I would like to see other options exhausted before resorting to such theoretical inelegance, and secondly, if we must resort to a plural view, it should be one that resolves rather than ignores the problems with its central components.

Instantiation

Instantiation claims that SE functions constitute a subset of CR functions. Griffiths (1993) and Walsh and Ariew (1996) each point out that if we consider an organism to be a system and if we consider survival and reproduction to be a systemic capacity of the organism, then we have a straightforward rule by which to consider SE functions to be contributing to that capacity, and thus to be a simple subset of CR functions. On this analysis, all functions are to be identified as causal-role functions, but the functions of biological traits are a special kind of CR function, which, thanks to natural selection's providing of a particular context for their systemic capacity, happen to owe their existence to the selective evolutionary process that produced them. Of course, the instantiation account really casts aside the SE analysis by differently accounting for the functions it had previously accounted for. This makes selection irrelevant to function—it is only the contribution to survival and reproduction, and not the selective history, that supposedly grants a function to an item.

It should be clear that the concerns with both the CR analysis and the survival and reproduction (SR) subclass of the RD analysis presented earlier still apply to this kind of a claim, whether or not SE functions are a subset of CR functions. For instance, the instantiation view

[250] Remember, logical conjunction (the "or" operation) includes things that are A, things that are B, and things that are both A and B; it only rejects things that are neither A nor B. It might seem convenient to say that, where the CR analysis fails, SE can step in, and where SE fails, CR can step in. But this would be an *unprincipled* alternation that gains us no theoretical ground and instead obscures the faults of both theories.

cannot account for sterile animals and for whole artifacts that are not, themselves, parts, nor does it resolve the innumerable counterexamples to the CR analysis catalogued earlier.

Let's Fix Theories, Instead of Kluging them Together

While the SE or SE-based plural views are widely touted as "consensus" (Allen and Bekoff 1995a; Buller 1999; Godfrey-Smith 1993; Neander 1991), the disputes and counter-intuitions published in a constant stream of papers and monographs show that very few are satisfied with any analysis that has yet been presented and that the keepers of the consensus even find little agreement amongst themselves as to the fine details of their variations. Buller (1999: pp. 19–27) summarizes the disagreement carefully, in an effort to highlight the so-called consensus. He suggests, "resolution of these [disagreements] constitutes the major part of the agenda for continued philosophical work on the biological concept of function" (1999: pp. 26–27). I am inclined instead to read the same level of disagreement as a possibly irresolvable philosophical confusion, based largely in the illusion of function constancy and the metaphysically unsustainable belief that *history*—natural selection—could play any role in the *current* causal circumstances of an item's functioning (see my discussion of "residues" above, on pp. 300–301).

E. Goal Contributions

To motivate a goal-contribution analysis, one may first observe that the functions of artifacts seem obviously determined by human goals. . . . Goal directed behavior also seems ubiquitous among living things, not just human beings.

—Christopher Boorse (2002:68)

The majority of thinkers about function and teleology have offered one variation or another of the three popular theories or their hybrids that we've just discussed, but their work certainly does not exhaust the thinking that has been done on these topics. It is time now to review the three less popular—indeed, I think, maverick—theories that will serve as some of the crucial building blocks for the theory presented later. I'll begin with Christopher Boorse's goal-contribution account.

On this view, if an item in any way causally contributes to the attainment of some goal for some agent, then the item is functional for the owner of that goal. There are a number of things that are vague about this presentation, not the least of which are the facts that Boorse does not specify any account of causation and that he does not give a theory of what constitutes an agent that could have a goal. However, one of the virtues of the account is Boorse's attempt at generality, by including both what he calls "weak" and "strong" function statements, which align respectively with the ideas of "serves a function" and "has a function". For Boorse, an item can *be functional* by contributing to a goal, but it can also *have a function* by being the kind of thing that contributes to a goal.

We can tabulate five general cases of how to apply Boorse's theory.

Case 1: *Biological Traits*. The functions of biological traits are all given by the goal-contribution (GC) analysis because of their contributions to the trait-owners' biological goals of survival and reproduction—we are talking here about things such as hearts, eyes, wings, gills, lungs, and opposable thumbs, as well as species-typical behaviors such as reflexive blinking or the burying of eggs by turtles; but not vestiges, such as the vermiform appendix or cave-dwelling creatures' degenerate eyes, and not spandrels, such as the wrinkle made between your bicep and your forearm when you bend your elbow. Spandrels don't count because, by definition, they don't contribute to anything; if they did, they would be traits, not spandrels (Gould and Lewontin 1979). Vestiges similarly don't count because, as vestiges, they no longer contribute to the goals their ancestral versions once did. However, during the interim period in which a species' changing relationship with its environment is transforming a trait to its vestigial form, the intermediary form of the trait would nonetheless have had a function, on Boorse's view, as long as it was still able to contribute somewhat to the organism's survival and reproduction. Such biological counterexamples as we saw before, in which a trait is seen as being somehow deleterious (including the liver's housing flukes and the narwhal's tusk dragging it down), are correctly counted out by the GC analysis precisely for the reason that these are ways in which they fail to contribute to survival and reproduction (though the liver's role in metabolism, bile production, detoxification and so on, and the ability of the narwhal's tusk to do whatever it does[251] would still be functions, and, of course, in terms of housing flukes the liver may still be functional . . . for the flukes). Our base case of the heart's pumping blood is correctly counted in since, by pumping, the heart contributes to the survival and reproduction of its bearer. In the case of an instant lion, Boorse's theory would not differentiate between the imaginary creature and a naturally evolved one, and would thus assign the functions that our intuitions would

[251] It is still not fully understood what the tusk's functions are but, as it is primarily a male trait of unusual proportions and also very dense with nerve endings the best current hypotheses are that it is a trait used in sexual selection (much like the peacock's tail) or as a sensory organ, or both. (Nweeia *et al.* 2014).

to any traits that helped the lion survive or reproduce. In the treatment of all these biological examples, there is a straightforward similarity between Boorse's GC analysis and the survival-and-reproduction (SR) analysis that I considered to be a subcategory of the RD analysis. Boorse remarks: "Since evolution in fact seems to yield organisms with the supreme goals of individual survival and reproduction (loosely, 'fitness'), within biology the [GC] analysis gives the same results as one defining biological functions, specifically, as causal contributions to fitness" (2002). The difference between these two analyses, as we'll see in just a moment, is that the GC analysis is also able to account for the functions of acquired behaviors, artifacts, and the traits of sterile animals, while, as we noted before, the SR and other RD analyses cannot.

Case 2: *Acquired Behaviors.* The functions of acquired, volitional behaviors, such as cooking, sailing, or playing squash are given by the GC analysis on account of their contributions to the intentional, psychological goals of their performer. Someone performs these behaviors because they want to engage in these processes or achieve their results. This interpretation even seems to be effective at accounting for biologically counterproductive behaviors such as suicide and celibacy because, as long as the performer intends the action, regardless of what might be *biologically* good for them, their *psychological* goal underpins the function. It is not yet clear, on an account such as Boorse's, how we should interpret psychological goals, nor why they would have the power to override the biological goals of survival and reproduction, but if we can decipher the relationship between psychological and biological goals, then Boorse's account—a contribution to any goal—will be satisfyingly inclusive.

Case 3: *Artifacts.* The function of an artifact—a hammer, a pen, or a chair—gets its status by contributing to whatever the user intends of it. The GC analysis can claim the function of a

hammer to be setting nails into place just in case someone uses it towards that end. A hammer might also serve the function of being a doorstop, a weapon, or a leverage fulcrum as long as someone chooses to use it in any of these manners to fulfill a goal. Our base case of a cog in a machine has the function of transferring mechanical force, on the GC analysis, since that is one thing it does and, in doing so, it contributes to a goal required by the machine's operator (say, the proper operation of the machine). If the cog accidentally falls into a machine and therein comes to be turned, although to no useful effect, then it does not function. But if, as Kitcher (1993) imagines, it falls into a machine and there comes to play a vital role in the operation of the machine, then despite the accidental nature of its functioning, it is nonetheless functioning. Our other base case, a rock, chosen as a paperweight, serves the function of holding down paper if it actually holds down paper, which is an effect that contributes to the goals of someone who prefers not to have to chase their sheets of paper around the room.

Case 4: *Non-Functions*. Our intuitions in the previously reviewed non-function counterexamples are upheld too. A cloud usually neither has the function of creating a rainbow nor of creating rain, since clouds have no goals; however, in some circumstances, the cloud can contribute to a person's goals, in which case it may function. For instance, one might suggest that a cloud may function if someone has the goal of getting another person to talk about light refraction and the cloud helps them accomplish that goal by providing a rainbow upon which to base the discussion[252]. As it turns out, Boorse's account already makes provisions for this: the cloud serves (but does not have) a function, for that person[253]. Similarly, the hole in the rubber hose that carries gaseous chlorine and

[252] This example is thanks to Colin Allen (personal communication).

[253] There is a more rarefied question as to whether the cloud functions if someone has the goal merely of having a rainbow appear. In one sense, the cloud contributes to the appearance of the rainbow and thus the fulfillment of that goal, but, at the same time, we might be unwilling to say the cloud functioned in this case, since the rainbow would have

Bedau's stick that is lodged behind a rock in a stream don't seem to be the kind of things that can or do have goals, and so Boorse does not assign them any function, but if they did support a person's goal of any kind, then they would serve a function, or "function as" something or other[254].

Case 5: *Accidents.* While accidental contributions count, such as the bible functioning to prevent a bullet from entering a soldier's heart, mere effects do not. If the bible prevented the bullet from entering a pile of sand we would not consider it having functioned to do so; it just would have happened to do so. This case is a possibly divisive one, though it shouldn't be. A proponent of the GC analysis would be satisfied that we have more broadly included more cases of (verb-) functioning without admitting any additional non-functioning, and would further point out that the popular analyses are unable to make this distinction: each of the CR, SE, and RD views would not consider either event to be an instance of the bible functioning[255]. In addition, the proponent would be satisfied that the GC analysis also includes other "weak" function statement cases, such as a rock used as a paperweight or a doorstop (FS7) or a wood box serving as a dog's sleeping quarters (FS13). However, an opponent who is convinced that the function–accident distinction is a central measure for theories of function might suggest that inclusion of the bible saving the soldier's heart is a case that shows the GC analysis casts too wide a net. Earlier I argued that the function–accident distinction is flawed for this very reason. On my view, what Boorse calls "strong" function statements—claims about items having functions—are just subjective perceptions based in the

---

appeared whether or not the goal existed. I think the proper interpretation of such a situation is that the cloud is making an accidental contribution, but a contribution nonetheless.

[254] Douglas Hofstadter (personal communication) has pointed out to me that the stick in the stream has served a number of philosophers' goals of finding a good example with which to discuss the notion of function.

[255] The RD and SE analyses would find the bible's performance in this manner neither to be a result of selection nor to have a propensity to contribute to replication. The CR analysis would rule, as it does with the stone paperweight, that there is no containing system nor overall capacity that the bible is contributing to in either the case of the heart or the sand pile—the bible is not a "part", it is just a whole object.

illusion of function constancy . . . all that a theory really needs to account for are Boorse's weak function statements.

Criticisms

A number of criticisms have been raised against the GC analysis. Boorse (2002) enumerated twelve possible concerns and he defended against ten, leaving two unresolved. I am going to give a blend of my defenses and Boorse's, here; however, a few of the issues will have to be revisited again later, in terms of the theory presented in Part II.

Objection 1: *Arbitrary, evaluative, or circular goal-choice.* Melander (1997) and Schaffner (1993) argue that making function relative to goals is only a vague, arbitrary theory. Melander (1997), for instance, notes that if one alters an animal's physiology so that its homeostatic systems approach different values than they normally do (say, so that the kidneys of a human maintain blood water content at 70 percent rather than 90 percent) there would still be a goal, and the GC analysis would have to assign a function. I think this is an important concern, and I am not completely satisfied with Boorse's response (but there's no need to repeat it here); Melander's concern can be resolved, however, by providing a different theory of goals—one that is not merely a cybernetic system with an arbitrary set-point—to which functioning can be made relative.

Objection 2: *Lack of explanatory power.* Neander (1983:98-100) and Melander (1997:36-8, 56) argue that a GC account of function does not allow a function to explain a trait's presence, which is, they claim, what function statements are meant to do (see also Wright 1973, 1976; Price 1995). I (and Boorse) take these thinkers' premise to be faulty. We can speak of present functioning without any

interest in the history of how an item came to be there (see also Amundson and Lauder 1994). We might also, for instance, be interested in *how the trait works* (Cummins 2002). The inference from a trait's functioning to its presence is both compelling and useful, as well as effortless to make once one is trained in Darwinian thinking, but that inference *begins* with a claim about modern functioning, proceeds via a number of logical steps that extrapolate that functioning back into history, assumes that the ancestral environment did not differ significantly from the modern one, and *ends* with a claim that establishes the trait's presence due to its ancestral successes. A function statement is part of the chain of reasoning that can explain a trait's presence but that explanation is not a direct corollary of the statement, and it is neither what the statement is necessarily meant to do, nor the only thing it can do.

Objection 3: *Artifact functions*. Nissen (1997) argues that Adams' (1979) account (which resembles Boorse's, 1976, version) cannot neatly deal with artifact functions because it requires outlining a system that includes both the artifact and its user, and then assigning a goal to that unusual and arbitrary system (such a view goes much more smoothly with organisms and their traits). This looks a lot like the system-bounding problem we faced when reviewing the CR analysis, which, in that instance, is a major concern. But Boorse (2002) sidesteps the concern by pointing out that, for his GC analysis, there is no need to look at such joint systems; the premise of the analysis claims that functioning exists when there is an agent whose goal the artifact contributes to. The goal is the agent's alone, not one of some larger system. I am inclined to agree.

Objection 4: *Environmental relativity*. This criticism was raised against Ruse's (1972) and Prior's (1985) SR and CR analyses, though not yet against the GC analysis, but Boorse notes that one day it could be used against him. The concern is that we may assign a function to a trait of an organism

only to see it disappear into thin air when the environment changes. For instance, polar bears have white fur, which has the function of camouflage, allowing them to hunt more effectively, but that function only exists in an environment of snow and ice (Melander 1997; Munson 1972). The SE theorist would like camouflage to be a permanent proper function of the whiteness of the polar bear's fur, until selection in the new environment changes it. This issue only affects analyses that intend to account for proper functions. Boorse tried to support both weak and strong function statements in his work, but my view of his theory is that it need not make (ontological) provisions for strong function statements. If there are no proper functions, if functioning itself is contextually relative as I have argued, then Melander's and Munson's example simply becomes a typical example of any functioning, and the concern simply dissolves.

Objection 5: *Maladaptive functions.* Neander (1983:89) charges that the GC account cannot distinguish between malfunctions and species-typical maladaptive functions. Boorse argues that, we shouldn't accept Neander's intuitions about the existence of maladaptive functions: if something is maladaptive then it is not a function. If the polar bear's environment were to become warm and snowless (as is now occurring) the whiteness and heat-retention of its fur would not be maladaptive functions, they would simply no longer be functions; they would be vestiges. I agree, but I would phrase it, only slightly differently, that these maladaptive features would no longer be functional (since TDHF). Neander is falling victim to both the malfunction fallacy and the illusion of function constancy. Maladaptive and malfunctioning items are both simply items that don't function; there is no need for a theory of functioning to distinguish between them.

Objection 6: *Functions vs. Accidents.* Neander (1983) argues that the GC analysis cannot make the judgments required by the function–accident distinction. As I've argued, the function–accident

distinction is a misleading result of the illusion of function constancy and is not a standard that an account of functioning needs to be held accountable to. As a matter of fact, a virtue of the GC analysis is that it nicely accounts for the accidental functioning of many items that other views, including Neander's, cannot account for.

Objection 7: *Unperformed functions*. Boorse describes unperformed functions as a generalization over malfunctions (things that supposedly have functions but cannot perform them) and unused functions (things that supposedly have functions, but in which, for instance, the initiating button is simply never pressed). The worry presented by Millikan (1989), Neander (1991), and Nissen (1997) is that the GC account can't attribute functions to either category of things, since such items never contribute to any goals. I think they are right—the GC account doesn't attribute functioning in these cases—but I also think that judgment is correct. Neither malfunctioning items nor unused items have functions (TDHF), but, more to the point, they don't even serve functions though, in some cases, they could. As with objection 5, this concern, too, dissolves in light of the malfunction fallacy and the illusion of function constancy.

I am going to jump ahead briefly to Objection 10, as listed by Boorse, for reasons that will become clear shortly after.

Objection 10: *Batesian mimicry*. Batesian mimicry is usually described using the example of the Monarch and Viceroy butterflies both of who have a similarly stark pattern of bright colorations, though only Monarchs have a taste that is unpalatable (to the birds that like to eat butterflies). The explanation for their similar appearance is that the Viceroy is an evolutionary mimic—its coloration pattern evolved to match that of the Monarch so that birds would avoid eating it, as if it were a

bitter Monarch. According to Mitchell (1995) the function of the Monarch's coloration is "to warn the predator of its unpalatability", while the function of the Viceroy's coloration is "to mimic the model and deceive the predator into presuming it is unpalatable." These different functions, Mitchell argues, can be determined by an SE analysis, which, it is claimed, uses the selective history of the traits in order to assign their functions[256], but, they supposedly cannot be distinguished by a RD analysis, nor (Boorse extrapolates) a GC analysis, which presumes that if the coloration plays a role in the goals of survival and reproduction, then both species of butterfly are equally served by the function, which can be said to be "to avoid predation". Boorse argues that Mitchell's conclusion is a simple mistake of underspecifying the mechanism and the function: For a GC analysis to say that something has a function (or, as I prefer to say, is functional) one need only suggest that it plays a role in a high-level goal such as survival and reproduction; but to say further just what that function is only requires a specification of the mechanism by which the goal is contributed to. Boorse says:

> In Monarchs, the function of the coloration does not depend upon traits of another species. In Viceroys, it does, namely, upon the poisonousness of Monarchs. It is this *present* difference between the two mechanisms, not its presumed evolutionary history, that Mitchell is describing when she says that the function of Monarch coloration is to warn, the function of Viceroy coloration to deceive. (Boorse 2002:104, emphasis added).

[256] This claim is highly suspect however, since one cannot determine the selective history of a trait without some clue as to the present (and then, by extrapolation, past) function of the trait (see also Amundson and Lauder 1994; Griffiths 2009; Pigliucci and Kaplan 2006).

The GC analysis can in fact determine the differing functions of Monarch and Viceroy colorations the same way biologists did in the first place—not by knowing their evolutionary history (which is something the biologists inferred *after* determining their current functioning) but by observing the characteristics of the two species, and the current relationships between the two species and between each of them and their predators.

Objections 8, 9 and 11 all relate to Boorse's statistical definition of normal function. He defines "medical normality as 'the readiness of each internal part to perform all its normal functions on typical occasions with at least typical efficiency'" and the term "typical" here is derived from an average over a reference class—the norm of a population "of uniform functional design: specifically, an age group of a sex of a species" (Boorse 2002; see also Boorse 1977). This helps Boorse make sense of strong function statements, which doctors need to make use of in practical medical contexts (in healthy individuals under rest conditions, the function of the heart is to pump blood with a pressure, X, at a rate, Y, where X and Y are distributions measured from a population and then used to diagnose whether a given patient's heart is functioning properly). These strong function statements are also a direct correlate of the proper functions that other theorists are interested in, which is why Boorse's account of them has drawn fire from those quarters. My response to this entire issue is simply that a theory of teleology need not directly account for proper functions. TDHF. Boorse should have explicitly separated his theory of functioning from his theory of function and considered them two different theories accounting for two different subjects. While biomedical normality is an important concept in medicine, and Boorse's statistical definition of that concept may be the most practical way to operationalize it, I think it is nonetheless a heuristic concept, and not useful as a basis for our philosophical theory of the fundamental nature of biology and function. After all, there are situations wherein the notion of functioning may apply but neither

having a function nor biomedical normality would be relevant. If an individual were to be born with a series of mutations such that (i) they have much thinner vasculature than normal (and thus a lower vascular volume, lower safe operating pressure for their veins and arteries, but a higher demand for flow in order to maintain perfusion of their bodily tissues), and (ii) in compensation, their heart pumped far more quickly yet with smaller, lower-pressure beats, such that altogether the individual was equally healthy as the rest of us, then biomedical normality would simply not apply to this individual. There would be no disease. And there would be no reference class to which we could compare the functioning of their parts. Still those parts would be functional (on a GC view as well as an SR or a CR view). Being functional is different from having a function. The latter is not a real part of the world (TDHF) but rather a heuristic that allows categorization of individuals based on useful similarities. I see no need to defend normal function or biomedical normality for my purposes; I take it that these concepts are practical extrapolations from the underlying facts of *functioning*, used to make useful comparisons but nothing more, and so I will set these issues aside until my new theory is laid out and we are ready to analyze its implications.

At this point, however, I must note one facet of Boorse's philosophy that is not in line with my own. As with the first three major analyses of function that we looked at, Boorse's version of the GC analysis turns out to be a non-normative theory. Boorse thinks that our analyses of health and health-related concepts need to be "value-free" because to be a "normativist about science in general, or biology, or biological function" is too "high a price to pay for [. . .] normativism about health" (Boorse 1997, p. 99).[257] In other words, Boorse seems to recognize the evaluative-normative nature of health and health-related terms but, motivated by strong materialist convictions, he is unwilling to accept that normativity as reflecting anything more than illusion. I disagree; I think the

[257] Boorse says elsewhere, "Clearly, by the biostatistical theory, insofar as biological function statements are normative, health is normative too. Once again I leave this issue to nonmedical philosophy of science. I merely note that for anyone who seeks to exclude values from scientific knowledge, holding biological function statements normative is a case of the high cost of normativism about health" (Boorse 1997, p. 58).

normativity we observe in medical science is fundamentally based in the normative nature of biology and, in particular, in the evaluative-normative nature of goal-directedness. We have only to account for that normativity.

Objection 12 is the most difficult of the bunch for Boorse to answer, and, presumably for this reason, he leaves both it and Objection 11 unanswered. Boorse describes Objection 12 as "attacks on the cybernetic analysis of goal-directedness" (2002), but I'd like to slightly reframe the issue as one of vagueness: The only naturalistic theory of goals given in the past century—the cybernetic theory—has been highly criticized and appears to be unsustainable in its current form (Bedau 1992b; Nissen 1997; Melander 1997). Without this underlying theory of goals, the GC theory of function is left with two of its major terms poorly defined. First, we are simply unable to identify natural "goals" in the world to which we can relativize GC functions and, second, it is unclear what it means for an event or a behavior to be a "contribution"—we are not prepared to identify the relationship by which a function may be relativized to a goal. Both "goals" and "contribution" are central terms of Boorse's theory, and yet both are undefined. In addition, if we were to have definitions for goals and their contributions, we might still ask what makes, say, (i) the contribution of an artifact to the goals of a user the same as (ii) the contribution of a trait to the goals of its bearer or (iii) the contribution of a behavior to the goals of its performer. What prevents the linking of these three (possibly) diverse types of contribution from being what philosophers call a category mistake?[258] Can a theory be developed that shows how all three of these are indeed, in some relevant sense, the same thing? If Boorse's theory (or my more normatively-flavored version of it) is to be sustained, all of these questions will need to be answered.

[258] A category mistake is simply when one co-categorizes things that have only superficial resemblance, and then tries to treat them as deeply similar.

Strengths

Aside from the GC analysis being so patently undefined, one of its strengths is that it is, compared with any of the more popular theories, better able to account for the various examples and counterexamples of function attributions that we make. While Boorse doesn't have a theory of goals, he shows that *if* we can come up with a theory of goals that is sustainable and that is also detailed enough to show how a goal can be contributed to, then we can probably support a clear theory of the ways we use the word "function", given in terms of those goals. While this conditional status of his theory appears to be a weakness because theories of goals are in short supply and there is widespread skepticism that one could ever be produced (see Chapter VII), I consider the same state of affairs to be a strength perhaps because I have more hope than do others that we can develop a theory of goals and also because many things about functions, function statements, and the counterexamples that we've seen all fall into place nicely if we do.

Another strength of the GC analysis, as I see it (but perhaps Boorse does not), is that it is inherently subjective—if goals are relativized to an agent and functions rest upon goals, then function is a subjective notion. Again, some might be concerned that this is a weakness rather than a strength, since they might want function to be an objective fact about items in the world. I don't. And I think those people might also be less concerned if it turned out we had a subjective view of functioning but could show how an objective underlying theory of goal-directedness accounted for the very existence both of subjectivity and of functioning.

F. Programmed Effects

A teleonomic process or behavior is one which owes its goal-directedness to the operation of a program.

—Ernst Mayr (1974/1988)

The second maverick theory we need to get acquainted with is Mayr's (1961, 1974/1988, 1992, 2002) teleonomic theory[259]. The word "teleonomy" was coined by Colin Pittendrigh (1958), and then adopted and further popularized by Mayr (1965, 1974/1988), Williams (1966) and Monod (1970), to describe "seemingly" goal-directed behaviors without admitting the existence of actual final causes in the world—it can be considered today to be about equivalent to putting the word "teleology" in scare-quotes. Mayr was interested in describing what he called "seemingly goal-directed behavior" (1974/1988).

Anything that is teleonomic, says Mayr, owes that status to its being directed by a *program*. Although Mayr's theory claims to be about "teleonomic processes in living nature" (1974), still, if we take it that functions are teleological (and if we ignore the scare-quotes implied by the term "teleonomy"), then application of Mayr's theory to the notion of function is fairly straightforward[260]: Mayr's view is that *functions are programmed effects* (PE).

[259] Cummins (1983), perhaps influenced by Mayr, also used the word "program" to elucidate his idea of functional analysis: "By a functional analysis, I mean an analysis of a capacity of a system into sub-capacities of that system such that exercise of the analyzed capacity is reduced to programmed exercise of the analyzing sub-capacities. By 'programmed' I simply mean organized in a way that could be specified in a program or flow chart. . . ." (Cummins 1983:29).

[260] Mayr (1992) expressly avoids the term "function" in describing his teleonomic theory. He borrows a distinction from Bock and von Wahlert (1969) between function and biological role. For these thinkers, the concept of *function* is more like a causal role (see also Amundson and Lauder 1994) while a *biological role* is more like what many other theorists would simply call "function". However, most theorists don't make this distinction, and so what is commonly called "function" aligns very much with Mayr's teleonomic theory of biological roles.

In what may be the most refined form given to the theory, Mayr wrote, of a program, "tentatively, *program* might be defined as *coded or prearranged information that controls a process (or behavior) leading it toward a given end"* and that endpoint should be "foreseen in the program that regulates the behavior" (1974/1988, emphasis original). While my review of the PE analysis in this section will amount primarily to unflattering criticism, I'd like to request the reader not to discard it from memory for its poor performance in its current form. I think there is a very important and widely overlooked core to Mayr's notion of a program as "controlling a process" that is worth preserving.

The notion of a "program" worked nicely for Mayr at the time of his writing because he was taking the analogy between computer code and genetic code rather seriously, and because the disanalogies between the two were perhaps not recognized as clearly as they are today. The notion of a "genetic program" seemed to allow categorization of all traits of biological organisms as teleonomic and Mayr's way of conceiving the design and building of artifacts as "programming" allowed the PE analysis to account for artifacts as well (Mayr 1974/1988). Unfortunately, Mayr later disavowed the inclusion of artifacts, saying they do not actually have programs in his sense and are only analogous to genetic and computer programs (Mayr 1992).

As an early example of how to interpret the theory (before his change of heart about artifacts), Mayr said, "The simplest program is perhaps the weight inserted into loaded dice or attached to a 'fixed' number wheel so that they are likely to come to rest at a given number. A clock is constructed and programmed in such a way as to strike at the full hour[261]" (1974/1988). In the case of the loaded dice, we can take it that the function of the weight in them is to shift their centers of gravity, causing the dice to more frequently land with the weighted side down and thus a certain desired number up—Mayr wants us to consider the foresighted placement of the weight to be the authoring of a mechanical program.

[261] Mayr was likely speaking of mechanical clocks not digital clocks, the latter of which are more literally programmed.

We can apply the view, only hazily, to our base cases. First, the PE analysis gives the function of the heart as pumping blood because, in some sense, the genetic code is the program that both built and operates the heart, causing it to pump. This success should be tempered, though, since the program in the genetic code also causes the heart to create heart sounds, and the theory is unable to differentiate between the results of pumping and these heart sounds. The theory is quite similarly unable to differentiate between such nonfunctional things as evolutionary vestiges and nonaptations or spandrels and actual functional traits, such as hearts, since all are genetically programmed in the same sense.

Second, a cog in a machine can be said to "have" the function of transferring mechanical force, on the PE analysis, because the design and construction of the machine, in having located the cog precisely where it is, is considered to be the process of programming that ensures the cog's behavior in doing what it does and in thereby contributing to the machine's overall behavior. If Mayr hadn't retracted his claim that artifacts have programs, this might have been the most convincing of his theory's applications to our base cases, but apparently he lost his conviction that mechanical "programming" had something fundamental in common with genetic and computational programming.

Third, the programmed effects analysis struggles to describe the function of a rock, chosen as a paperweight; it seems that we would need to significantly broaden our notion of programming if the mere—perhaps even careless—selection of a rock and placement of it upon a stack of papers is to count as a program. In all three base cases, it appears that a clearer definition of what a program is would be required in order make sense of the claim that everything functional is directed by a program.

Criticisms

In his review of the PE analysis, Nagel (1977) gives two counterexamples to Mayr's view. The first is a biological example that seems by all lights to be a Mayrian program: the patellar reflex (Nagel calls it the "knee-jerk reflex"), which Nagel claims is not goal-directed and which has no obvious function. The second counterexample is radioactive decay, which, Nagel points out, occurs in a manner that seems to mechanistically proceed towards a predetermined end, much like the clockwork that Mayr suggests includes human-made clocks as being teleonomic. No one would want to include radioactive decay as being a teleonomic process, of course. Both examples are excessive inclusions of the theory, and so Nagel suggests that maybe Mayr's idea needs to be narrowed to programs "of a *special kind*".

Nagel's second counterexample incited a response from Mayr in a postscript to his 1974 paper added in 1988, but the response is difficult to interpret. Mayr wrote "radioactive decay is controlled by laws and not by any particular program; it obeys the same laws any time anywhere. Programs are highly specific and often unique." It is not clear what distinction Mayr is making here—is he suggesting that radioactive decay fails to count because it lacks conditional branching? Is uniqueness or rarity in a process an additional constraint for being teleonomic, and how would we determine it, if it were? Mayr (1992) remarks that "information and instruction" are the key notions to what makes a program a program, and this does seem to provide a principle which could exclude Nagel's counterexample, but it also may require a reanalysis of whether—or in what sense—artifacts such as weighted dice or a mechanical alarm clock involve information and instruction.

Allen and Bekoff (1995), channeling a personal communication between themselves and Elliott Sober, make this last criticism more explicit.

> Mayr's use of the notion of program here is an unexplained metaphor. For example, where is the program in a clock? Sober objects that the idea of a clock being programmed simply amounts to the claim that someone designed it. We agree that Mayr's use of this notion does not conform to the literal sense in which computers may be said to execute programs and that Mayr must therefore explain his use of the notion before his account of biological teleology can be accepted. (Allen and Bekoff 1995a)

Allen and Bekoff (and Sober) want to know what allows us to extend the idea of a computer program to organisms and artifacts. In what sense do these all have programs? In what sense do their performances all correspond to the orderly execution of an instruction set?

There may not be an answer to satisfy these questions but, despite the counterexamples and vagueness, I think there is something important in Mayr's notion of a program causing functional behaviors. His intuition that everything that is purposeful seems *somehow* programmed, directed, or instructed is insightful and should be added to our list of intuitions and explanatory desiderata.

G. Valuable Effects

The biologist who helps himself even to such an obviously safe functional category as eye*,* leg*, or* lung *is already committed to assumptions about what is good.*

—Daniel Dennett (1987:278)

Our third maverick theory is what I will call the valuable effects (VE) analysis. In the title of a standout paper responding to the lifelessness of CR, SE, and RD theories, Bedau wonders, "Where's the good in teleology?" (1992b). The same general sentiment—that good, or value plays a role in teleology—has also been called "the welfare view" (Nagel 1977) or the "good consequences" view (van Parijs 1981). It has various precursors in Hempel (1959/1965), Sorabji (1964), Elster (1979) and the early SR analyses (*e.g.* Canfield 1964; Ruse 1971; see also Ayala 1970) all of which focused on what is useful or necessary (*i.e.* valuable) for an organism, under the assumption that survival and reproduction are the ultimate operating principles of organisms[262]. In recent decades, however, only a few writers have supported a theory of function in terms of value (Bedau 1990, 1991, 1992a, 1992b, 1996; McLaughlin 2001; van Parijs 1981) while most other modern writers have largely ignored it. The version that both van Parijs and Bedau support is roughly that *functions are the consequences produced by a trait, an act, or artifact that are good or valuable for either the bearer of the trait, the performer of the act, or the user of the artifact.*

Before we assess the merits and weaknesses of the VE analysis, it is necessary to review some of its nuances. In particular, Bedau's (1992b) version distinguishes between three grades of involvement of value in teleology that help to clarify how to apply it. The first grade refers only to

[262] To be sure, the idea of "the good" being relevant to teleology goes back as far as Aristotle who often referred to his final causes using the phrase "the end and the good".

the production of good consequences. This grade is very much like Boorse's GC analysis—it accepts the soldier's bible as being functional since, in stopping a bullet, the book produced a good consequence, for the soldier. However, Bedau differs from Boorse in that he would like his theory to exclude this kind of accidental effects, and so he discounts the first grade in favor of the next two, both of which can be seen as versions of the SE analysis modified with a value constraint.

Grade-two teleology, on Bedau's account, involves something that occurs *because* of its consequence, which happens to be good for someone, though *not* because that consequence is good. Bedau says, "In grade two teleology the consequences cannot be accidental, but the benefit they provide can" (1992b:789). The soldier's bible is now already excluded—the bible stopped a bullet and that happens to be good for the soldier, but its occurrence was accidental. The heart's function of pumping, on the other hand, fits the second grade straightforwardly: The behavior of pumping happens to be good for someone (the heart's owner) and also that pumping occurs non-accidentally, as in Wright's SE analysis, because the pumping of hearts contributes to the creation of hearts (through reproduction and selection).

Bedau's Grade-three teleology can be distinguished from grade two because "good consequences and their goodness both figure in the explanation" (1992b:790). This is what Bedau considers to be full-blooded teleology. To put it in other words, an item exhibits grade-three teleology if its consequences, which are good for someone, occur in part because they are good for someone. However, Bedau only considers mentally directed behaviors and artifacts to be archetypically grade three. For instance, we walk to the grocery or sit in a chair because we believe it to be good for us and so the goodness of the effect of either the act or the artifact plays a role in an explanation of the occurrence of that effect (while no such belief about the future plays a role in how the heart comes to be a pump). The heart and other biological traits are not grade three.

Bedau says "Since the goodness of survival does not itself play a role in natural selection, biological teleology never surpasses grade two teleology" (1992b:802).

If we measure Bedau's VE analysis up against our base cases we find that it performs rather well. The heart's pumping is good for its owner. The cog's performance in the machine has a good effect for an operator who gets some value from the machine's operation. And the stone paperweight is similarly good for the person using it to hold down their papers.[263] Of course, as noted, on Bedau's account, the heart is grade two while the cog and the paperweight are grade three, meaning that the heart does not have full-blooded teleology, but is only "a cousin of a central family member" (1992b:790). Bolstering this performance with our base cases is a lack of counterexamples in the philosophical literature, though it is not clear whether no one has found any or whether no one has tried. For my part, nothing obvious has come to mind. Boorse, noting that Bedau's second and third grades are descendants of the SE analysis, says "if functions are [in particular] *etiologically significant* contributions to value, then Bedau suffers all the same counterexamples of unselected function as a pure SE view" (Boorse 2002:68, emphasis added). Boorse is referring to such things as segregation distorter genes and clay crystals; however I am as unconvinced about using these examples against the VE analysis as I was when they were used against the SE analysis, particularly because I think our intuitions about whether these things are or are not functional may be unreliable. They are not clear cases.

So, as we've done with other theories, we can summarize the cases in terms of some categories: The functional traits of organisms all serve the good of that organism, as long as what we take to be the organism's "good" is hitched to its survival and reproduction. Vestiges, on the other hand, no longer produce any good of that sort and so are no longer considered functional.

[263] This all follows straightforwardly from Bedau's statement of applicability: "Where [the functional item] is an organ, the beneficiary is the organism containing [the organ], and where [the functional item] is an artifact, the beneficiary is the person using [the artifact]" (Bedau 1992:792).

Nonaptations never produce any good. The liver may function as an artifact for liver flukes, much the way a stone paperweight functions for a person, and while the behaviors of oncogenes may not be good for the afflicted organism, they can be seen as good for the cancer cells themselves. In contrast with the RD analysis, the VE analysis is able to account for the traits of sterile organisms, because, for Bedau, value is defined in terms of the individual, rather than in terms of the survival and reproduction of lineages (1992b:791) and so mules can be beneficiaries just as horses and donkeys can. In terms of artifacts, they are functional if they contribute to the good of an individual user, which is what most artifacts do, when used. There is an open question as to whether the artifact has to actually have contributed or whether it is sufficient to just be capable of contributing (think of a never-used item that has come off the assembly line and sat in a drawer ever since) though a proponent of the VE analysis could likely add a rider that makes sense of either perspective. Bedau, by distinguishing the constraints of his second and third grades of teleology from the first grade, has also carved out any item that accidentally contributes to an individual's good. In sum, the VE analysis seems to give a fairly sturdy performance in terms of examples and counterexamples.

Concerns

In some sense, the VE analysis may appear to differ very little from Boorse's goal contribution analysis, if one takes it that things that have value are precisely those things that contribute to goals (again, either psychological goals or survival and reproduction as biological goals). However it is not necessary that all conceptions of goals are value laden and, in fact, Boorse's own use of the cybernetic theory of goals can be seen as a value-free notion (cybernetic

goals are simply *feedback systems*). Boorse himself, unmoved by the idea of natural value, finds the resemblance unpersuasive.

> I am unconvinced by the efforts of Bedau and his sources, Taylor (1986) and Callicott (1989), to show that all living organisms have intrinsic value of a kind that artifacts such as watches or pianos do not. Writers who attribute intrinsic interests to plants or bacteria, or a good or welfare of their own, are, I would argue, either anthropomorphizing, or advocating incomprehensible values, or confusedly referring to some descriptively definable property such as life or goal-directedness, in which case the [VE analysis] metamorphoses into a [GC analysis]. (Boorse 2002)

My primary concern with the VE analysis is slightly different than Boorse's. I agree that the division between artifacts and organisms is made too sharp by Bedau's view but contrary to Boorse, I do think there's something to Bedau's (and Taylor's and Callicott's) notion of an "intrinsic value" in living organisms (see also McLaughlin 2001). The way those four authors have discussed that intrinsic value, though, is rather underspecified, and that brings me to my central criticism of the VE analysis: I find the notions of "value" and "good" as they are used in the analysis to be unworkably vague. Bedau does present an analysis of what he considers "X is good for Y" to mean. He says it amounts to:

(i) Y is the kind of thing that has its own interests (a "good of its own");

(ii) Y's good is independent of any value that some third party might place on Y; and

(iii) X is in the interest of Y, *i.e.*, X promotes Y's interests or constitutes (at least part of) Y's interests (Bedau 1992a).

But this analysis is still nebulous because the underlying notion of "interests" is also not well defined. If a person is interested in committing suicide, is such an act in their interests? It is hard to say, since the act may not be in their biological interests but it may be in their psychological or ideological interests. If interests can conflict like this then it is hard to know how to apply Bedau's analysis. What exactly does "interest" mean? Moreover, what criteria are we to use to determine whether an item is the type of item that has "a good of its own"? What kinds of items does it apply to, and how do we know? As Boorse pointed out, relying on the term "life" or "living" doesn't get us any closer since it is debated just what fits into that category too. It is likely that this imprecision is what has motivated other theorists to simply ignore Bedau's account much of the time.

McLaughlin (2001) attempted to define beneficiaries as self-reproducing systems[264]. On this view, then, value would be anything that is good for such a system—that helps it in its self-reproduction. I think it is a good start, but not quite general enough (McLaughlin arrives at his conclusion at the expense of casting aside reproducing—replicating—systems as unable to be beneficiaries).

[264] By which McLaughlin seems to mean, more or less, autopoietic systems, although he avoids that term and avoids citing any of the autopoiesis literature.

H. Convergence

Saying "An elephant is like this, an elephant is not like that!" [and] "An elephant is not like this, an elephant is like that!" they fought each other with their fists. And the king was delighted (with the spectacle).

—Udana 6.4 (Trans. John D. Ireland, 1997)

Our tour through the garden of function analyses is now finished. Along the way we looked at six primary theories: Causal Roles (CR), Selected Effects (SE), Replication Dispositions (RD), Goal Contributions (GC), Programmed Effects (PE), and Valuable Effects (VE), as well as a few variations of those more central analyses (SR, self-replication dispositions, and various types of pluralism).

We found that the three popular theories each fail to account for all the ways we would like to use function statements and for the intuitions about functioning that I developed in chapter IV. At the same time, though, all three seem to have some potential concessions that, were they to be made, could perhaps allow amendments that would make up the difference in their explanatory scope. On one hand, some of those concessions are large, and it is debatable whether amended versions would resemble the originals enough to be called descendants. For example, the SE and RD theories would have to relinquish natural selection, which forms the heart of both analyses and, at the same time, the CR theory would have to accept normative, teleological aspects of function, the rejection of which has always been a cornerstone of that analysis. As I said, these seem like enormous, insurmountable concessions to make, but that's what it would take to find a middle ground (aside from the pluralist attempt to simply agree to disagree). On the other hand, I think it is still the case that large parts of these theories will remain intact in a complete account. Among the

things I don't think we would want to discard are: The CR notion that functional items play causal roles in some kind of an organizational structure; Wright's idea of self-causation; Millikan's notion of direct and derived functioning; the general focus of the SE account upon normativity; and the RD analysis' focus upon contributions to reproduction or self-reproduction. If a new account can be found that comprises all these aspects in some manner, it may well be a happy medium that can make sense of the many ways we see function in our world.

As for the three unpopular theories, they each appear to have a very fundamental kernel of truth to them, but also a fundamental vagueness that undermines the theoretical power of that truth. If we can develop a meaningful theory of goals and what counts as a causal contribution to one (Boorse's GC analysis), a definition for a general-purpose concept of "programmed effects" (Mayr's PE analysis), or a naturalistic theory of what value or good consequences might be and who could be a beneficiary (Bedau's VE analysis), then it seems likely that one of these theories might also be able to explain the teleological notion of function. As it turns out, I think that the answers to all of these quandaries will converge and that when we understand them in their proper contexts, they will come to resemble one another. Moreover, I think that the CR, SE, and RD analyses, as I imagine them to be amended, will also converge upon the very same result. It is this convergent view that I intend to outline and advocate in the coming chapters. In the end, none of the analyses will be quite the same as it is at present, but each will have moved closer to the others and we will have held on to most of the useful core insights from all of them.

Chapter VI

Twenty Questions For a Naturalistic Theory of Teleology

Naturalism is the doctrine that everything real is at least in principle within the scope of a purely scientific account of the world.

—Mark Bedau (1991), writing on teleology

We've covered a lot of ground so far in our explorations. At this point, I'd like to give a brief recap of where we've been, and then to use that retrospective to help map out the rest of what's to come.

In the first two chapters, I began to make a case that the notion of goal-directedness (and it's subjective nature) is not just a common thread running through the domains of agency, biology, cognition, computation, and technology; it is the key pattern that accounts for the very existence of those diverse phenomena. We found that the problems of identity and value—what I called "the fundaments of subjectivity"—are perhaps the central unsolved riddles challenging us in bringing goal-directedness and other subjective topics under scientific consideration. In Chapter III, we reviewed much of the history of teleological and vitalist thinking, discovering many of the observational patterns that motivated early thinkers on these topics, and watching as these two streams of thought waxed and waned in varying quarters over the past few millennia. Then, in Chapters IV and V, I attempted to mine the most valuable insights from the recent biophilosophical work on the subject of function. At this point, Part I of my account is nearly complete; I will finish it off in the next chapter by making an appeal in favor of pursuing a theory—rather than a dismissal—of teleology.

While I think we've made good progress in reframing many of the questions that surround teleology, so far these explorations have left us with far more loose ends than resolutions. By the end of Part II, I intend to show how we can tie up a good number of those loose ends, but there is much work to do before we are ready to digest those proposals in full.

The first thing to do will be to set out the goal posts, and so I'll turn to that project now. In a previous piece of theoretical work, I used a list of twenty questions to frame the goals for that theory and then, later, to test whether those goals had been met[265]. I'll repeat the same strategy here. Below, I present a list of questions, in twenty principal categories that map out the explanatory ground that we would like a theory of teleology to cover. The first six are derived from issues raised in chapters I through III, the next seven represent the issues raised in Chapter IV, and the last seven summarize the issues raised by the theories described in Chapter V. My claim is that any theory worth its salt—whether it is the one I present or one developed from other considerations—should eventually lead to convincing answers for all twenty of them.

1. ***What is teleology?*** If subjective goals are a real part of our objective material world, then just how does that state of affairs come to pass? What are goals made of, and how are they materially realized? Quite simply: *What does it mean to be goal-directed?* Not all *ends* are goals, so which ones (if any) are, and why should that be so? Why are perseverance and plasticity widely thought to be the hallmarks of goal-directedness? And what is the relationship between biological goal-directedness and the more salient psychological teleology of conscious, representational beings such as humans?
2. ***What is identity?*** How can we carve an individual from its environment? Or distinguish it from other individuals? Is there a principle by which we can identify an agent that can be

[265] Hurley, Dennett and Adams (2011).

said to be the locus of goal-directedness—that is, when we say that something is goal-directed or purposive or functional, how do we define that "something" so that we know just whose goals or purposes we are talking about, or who is benefitting from an item's functioning?

3. ***What is value?*** Speaking of benefitting, is there also an objective way to account for the subjective notion of value? What types of items or patterns in the world is the notion of value relevant to? Can value accrue for just any item or object or is it relevant only to certain types of things? How and why? And what is the relationship between value and goal-directedness? Can we account for the evaluative norms by which the achievement of goals or the serving of purposes can be benchmarked?
4. ***How do we differentiate teleological patterns from other patterns?*** Both teleological patterns and spontaneously organizing patterns appear to metabolize free energy to build structure. How do they do that? How do we solve the material coordination problem: what is the source of information—the blueprint—that helps these kinds of open systems produce their orderliness? And what is it about teleological patterns that make their metabolic processes different from those of spontaneously organizing patterns?
5. ***What is life?*** What makes living things alive? What are animacy and agency and vitality and projectivity? Can a natural, material theory of teleology help restore older conceptions of life that depended on a more vague notion of purposiveness (as opposed to those more recent conceptions that cite 19th and 20th century biological concepts such as metabolism, reproduction, cell membranes, DNA, and evolution)? Moreover, can a material theory of teleology help revitalize vitalist notions, such as entelechies and élan vital, in an emergent and non-enchanted way? Can it describe the kind of organization that matters in the

biologist's now-common phrase "organization matters"? In just *what way* is a biological whole more than the sum of its parts?

6. ***How can we account for future-directedness?*** Goal-directedness is most certainly about the future, but how can a thing be "about the future", especially if it is not an intentional, psychological system? Can our account of biological teleology be made consistent with a conception of causation in terms of physical forces and Eddington's arrow of time, and still, in some way, be about the future?
7. ***What is function (and how does it relate to teleology)?*** Is it possible to fill in the blank, in the statement "the function of the heart is that particular thing the heart does that _____", in a way that is so general as to allow us to transplant "the heart" with any other functional item while maintaining the truth of the statement? Or if proper functions do not exist, then what is function*ing*? What particular type of relationship must exist in the world for us to say that an item is functioning? And just how does having a function (even if it is an illusion) relate to serving a function? Is function a necessarily teleological phenomenon? Certainly it isn't only goal-directed things that can function (a hammer isn't goal-directed), so then what exactly is the relationship between function and goal-directedness?
8. ***Can teleology provide for the autonomy of biology?*** Is function the fundamentally irreducible characteristic of biology that Laubichler suggested it is? Can a law-like regularity be used to account for all functioning? Or, if function is derivative (that is, if it is a relational property as I've claimed), then what relationships does it derive from and how might that phenomenon instead contribute to this apparent irreducibility? If there is some teleological law of biology, how does it relate to other suggestions for biological laws? What are we to make of Rosenberg's assertion that natural selection is the law that gives biology its

autonomy? Or McShea and Brandon's assertion that "biology's first law" is the tendency for diversity and complexity to increase? Might there be two or three (or more) laws of biology?

9. ***What accounts for the variety of function statements?*** Can a theory of functioning, whatever it is based in, account *in some way*—whether it is in theoretical, analogical, or cultural terms—for all our function statements? How do we make sense of some of the more exceptional cases such as the function of the ozone layer; or of free radicals in the ozone layer; or of rhizomes in the forest ecosystem; or of a non-designed rock used as a paperweight? Can our theory of functioning make sense of the intuition that there may be more than one function per trait or per part of an artifact? Can it account for the fact that purposiveness appears limited to organisms, their behaviors, and their artifacts?

10. ***Can the many senses of "for" be unified?*** When we ask the teleological question "what is it for?" there are many possible interpretations of what the word "for" means. I suggested that this could cause some confusion, but that each sense of "for" may ultimately be in some way derivative from the fact that items with purposes or functions are personally good for some agent. Can our theory of teleology either account for this suggestion, or else replace it with another explanation?

11. ***What is design?*** Why do things that are designed coincide so regularly with things that have functions? Even if not quite all things with functions are designed, why are natural design and human design both such prodigious progenitors of functions? Can our theory of teleology clarify the relationship between design and function? The relationship between design and agency? Just what is design?

12. ***Can we account for accidental functioning?*** Can our theory of functioning equally account for both intended functioning and accidental functioning? And if we can account for

accidental functioning, can we thereby avoid using the function-accident distinction to judge the theory of function? Is there another benchmark that we can use?

13. ***How can we reform our concept of function?*** Even if the concept of a proper function is an illusion, thinking of items as having functions can be a useful tool in doing functional-analysis and reverse-engineering in order to figure out how artifacts or organisms may work. If we develop a new theory of function, and if we discard the function-accident distinction, discredit the notion of "proper" functions, and discount intuitions based in design, malfunction, and items being for something-or-other . . . what remains of our concept of function? How much does the new concept resemble the old one? Can it still be used to perform the same theoretical jobs? Can it still refer to things the way it currently does in everyday usage?

14. ***Why does playing a causal role in a containing system correlate so well with items that function?*** Why are functional items commonly found to be parts of systems that are hierarchically structured? Said another way: Why is the process of functional analysis such a successful strategy when reasoning about both biological and artifactual items? Some causal roles are not functional—the sun's gravity tugs ever so slightly on our hairs but it doesn't "have the function" of making our hair just a smidgeon lighter in the day and heavier at night—so just what kind of causal capacity or structure is required for functioning? Does a theory of function need to provide a definition of a "system" or a "part" of one and, if so, how does our theory specify these things? Moreover, how does our theory of function make sense of an item such as a stone paperweight or a cup or a stick of chalk that plays a functional role without having parts or belonging to a system?

15. ***Why do the products of natural selection correlate so well with items that function?*** Why do most functional items seem to be either the products of natural selection

(organisms) or the products of the products of natural selection (artifacts and behaviors)? If natural selection doesn't create function then how else might we account for the regularity with which the two are associated? Can our theory account for Wright's intuition that the function of an item is whatever it does to cause itself to be there? Can it give a more unified account for both artifacts and organisms?

16. ***Why does membership in a reproductively established family correlate so well with items that function?*** Is membership in such a family required for functioning? If so, how shall we determine an item's membership—when is an item that is made by some parent item identical enough to be considered a child, and when is it not? If such membership is not required, then what else might account for the fact that the majority of functional things appear to be such members? Also, does our theory provide a principle by which we can not only distinguish functions from malfunctions, but also distinguish each of those categories from non-functions? Does the new theory account for Millikan's ideas about direct and derived proper functions? Does it account for items that are functional without being the products of selection? How does it classify doubles such as instant lions? And the sorting processes that Lewens describes as using the same norms that selection does? Most importantly, does it avoid the SE theorist's metaphysical dilemma of positing either a vital residue such as a function, or a causally irrelevant notion of function?

17. ***Why does the disposition to replicate correlate so well with items that function?*** Is there something about replication (even without selection) that is important to functioning? What is it about the biological notion of survival and reproduction that correlates well with items that function? Is the altered form of this—survival or reproduction—a useful notion despite appearing to be disjunctive? Can our theory somehow maintain this basis in replication and yet still account for the functioning in limited edition designs that don't have replicative

futures, such as artifacts and sterile animals? Can it also account for Bigelow and Pargetter's intuition that for an item to exist, it must have been serving its function all along?

18. ***Why does the contribution to a goal correlate so well with items that function?*** If functions aren't contributions to goals, then why else do functional items always seem to serve either an agent's psychological or its biological goals? And if functions are contributions to goals, then in what way is that relationship structured? What exactly does it mean to be a goal? And what does it mean for something to be a contribution to one? How does our theory compare with the cybernetic theory in buttressing Boorse's notion?

19. ***Why does the notion of programming correlate well with teleological items?*** What is it about organisms and artifacts that make both somehow seem to be programmed? Are information and instruction the central notions that define programs? If so, what would allow us to extend this instructional idea of a computer program also to organisms as well as artifacts other than computers? If not, what other pattern might account for Mayr's intuition? Is there a phenomenon that equally subsumes computer programs, artifact design or use, and the behavior of organismic traits in some way such that all may somehow be seen to be the same?

20. ***Why does the notion of value correlate so well with items that function?*** Everything that functions seems to provide value to an agent; but just what does that mean? And if functioning is not to be equated with having valuable effects, then what else might account for the two phenomena being so commonly associated with one another? Bedau talks vaguely about "the kind of thing that has its own interests (a 'good of its own')", but just what "kind of thing" is that? Is value somehow intrinsic to *organisms,* or is some other category more precise? If we know what kinds of things can accrue value, then do we also know how value may be conferred upon them? And, lastly, if value or interests turns out to

be a key term in a theory of goal-directedness, what sense can we make of conflicting interests, such as the suicide bomber's biological interest to live and their ideological interest to die—can both these things be good?

While this list may not be complete, it is certainly a good start. If a theory of teleology is held accountable to answer all the questions here, that theory will not only provide an explanation for the basic teleological notions of goal-directedness, value, and identity, but it will also elucidate many of the issues that make function a thorny topic and it may explanatorily subsume all the major families of function-theory from the past century.

Ideally, though, one would like a theory to do more than just that. No doubt, if teleology is as important a subject as I have claimed, then a proper theory of it will bear further upon subjectivity, agency, will, and the relationship between agents and technology as well as the search for non-biological life, and the development of artificial intelligence. Furthermore, a proper theory should be able to explanatorily replace previous theories in their extended roles, for instance in buoying theories of biosemantics, biomedical normality, functional analysis, and ethics. It should either leave existing teleologically based or function-based theories of these topics largely intact yet set them on a somewhat modified foundation, or it should expose weaknesses in them and offer ways to repair those weaknesses. Beyond this, we would hope the theory to eventually make new and interesting predictions in many or most of the related fields just discussed. Near the end of the dissertation, I'll very briefly and speculatively address all of these topics, leaving the bulk of those explorations to any theorist or experimentalists who find the proposals I've made interesting enough to pursue.

Chapter VII

Realism about Goals and Purposes

I have stressed the importance of the use of such concepts as biological means and ends because I want it clearly understood that I think that such a conceptual framework is the essence of the science of biology.

—George Williams (1966:pp)

Rather than reject this idea (as certain biologists have tried to do) it is indispensable to recognize that [goal-directedness] is essential to the very definition of living beings.

—Jacques Monod (1971:pp)

Purposefulness, or teleology, does not exist in nonliving nature. It is universal in the living world. It would make no sense to talk of the purpose or adaptation of stars, mountains, or the laws of physics. Adaptedness of living beings is too obvious to be overlooked.

—Theodosius Dobzhansky, *et al.* (1977)

Larry Wright opened his classic paper, "Functions", by noting the non-centrality of the concept. He wrote, "The notion of function is not all there is to teleology, although it is sometimes treated as though it were. Function is not even the central, or paradigm, teleological concept" (Wright 1973). The paradigm concept he is referring to is of course what we refer to variously as goal-directedness, striving, or purposiveness. Still, over forty years later, the bias that Wright

pointed out persists in the literature. It is exceedingly rare to find a paper that presents or defends a theory of goal-directedness while, as we've now seen, the debate on function continues to flourish.

The main reason for this theoretical gap is the challenge that goal-directedness has long presented to the materialist. There is nothing in the material world of particles, forces, and so on, that seems even remotely related to values, reasons, intentions, or strivings. Where would such subjective properties come from? The cybernetic approach—the only serious attempt to make objective, scientific sense of goal-directedness during the twentieth century—fell out of favor around Wright's time and, until recently, no new offerings have been made. The topic of function, despite many challenges of its own, has just seemed so much more approachable by comparison.

Standing in contrast to the indifference of the material world are two everyday experiences. The first is our own psychological experience in which every action we take seems to be motivated by our psychologically goal-directed nature, by our individual aims and ambitions. The second is our observation of the goal-directed nature of other organisms—of bacteria and sunflowers and mosquitoes and squirrels—striving in various ways toward various goals. And so we are trapped—caught between, on one hand, this distinct impression of biological and psychological entities and, on the other hand, our scientific understanding of the material world. We find ourselves convinced that spirited organisms are composed of spiritually inert matter, and yet unable to make sense of the lively former in terms of the lifeless latter (see also Dennett 2017).

In this chapter I will describe this problem, which I'll call the teleologist's dilemma, and I will compare it to a similar dilemma faced in the last century by behaviorist psychologists. I'll explain why I think the modern teleologist's focus on functions instead of goals stems from the same concerns that caused behaviorists to emphasize behavior rather than minds; and I'll suggest that the logical similarity in their motivations is why the teleologist's focus also, unsurprisingly, suffers from some of the same flaws that plagued behaviorists. Then, I will describe what I'll call

"eliminativism"—the commonly held view that appears to resolve the teleologist's dilemma by claiming that goal-directedness in organisms is only "apparent" or "seeming" (while leaving human psychological goal-directedness simply unexplained, though loosely assumed to be a product of how the brain produces a mind). I will cite a broad variety of scientists to show that holding some form or another of eliminativism is the fashion of today. Lastly, I aim to show that eliminativism is an unsustainable position and that we should instead believe in the goal-directedness that we all observe in the world until an enterprising eliminativist can clearly demonstrate how the illusion that they propose to exist is constructed, or under what types of limited conditions it occurs.

A. The Perception of Goal-Directedness

You cannot conceive of a living organism, not to speak of behavior and human society, without taking into account what variously and rather loosely is called adaptiveness, purposiveness, goal-seeking and the like.

—Ludwig von Bertalanffy (1968:45-46)

Nothing could be more obvious. Organisms have aims and purposes, which their behavior serves; their component parts serve to fulfill these purposes and have functions in meeting the needs of cells, tissues, organs, whole biological organisms, and systems like ant colonies made up of a large number of individual organisms.

—Alexander Rosenberg (1985:43)

When we look at biological organisms of any type, their goal-directedness is immediately apparent to us, from their structure to their behavior, at every level of their organization. As Rosenberg says: "Nothing could be more obvious."[266] For now, let's call this an observational or perceptual fact, rather than a biological fact. That is to say, I want to draw attention only to the fact that biological organisms *look* goal-directed (particularly in the sense of "goal-directed" that is not based in psychological goals). We can remind ourselves of the breadth of this perceptual fact by examining a handful of cases.

Think, for instance, of a ground squirrel. It spends its summers and autumns toiling away in order to collect calories it can use to survive through the winter. It does this in two ways: first, it is biochemically inclined to accumulate substantial deposits in its fat cells that will both insulate it and

[266] It is interesting to note that, after making this rhetorical claim, Rosenberg goes on to argue that the fact which is so obvious is, in actuality, an illusion (Rosenberg 1985, 2013). I will take his claim more seriously than he himself does.

slowly burn off during the colder months, and second, it is behaviorally inclined to sock away hundreds of seeds and nuts in carefully chosen (and remembered[267]) hiding places so that it can later eat them in order to replenish both nutrients and energy. The animal even has specialized parts and behaviors used to help achieve these goals. For instance, it has (non-salivary) pouches in its cheeks that it uses to cache more than a meal's worth of food so that it may free its forelegs for running off to "squirrel away" the nuts, as we say.

Figure 7.1: A hungry squirrel busy collecting nuts, in order to prepare for the winter.

In everything they do, from morning to night, squirrels appear to be goal-directed. Later, we'll run across one author who refused to believe in purposiveness in squirrels (Ducasse 1925). Of course that author would not have found squirrel goal-directedness worth even a momentary

[267] Jacobs and Liman (1991) observed that squirrels retrieve nuts from their own stashes, but not those from each other's stashes buried in the same area, showing that retrieval depends on more than just the smell of nuts.

mention if he had not clearly observed their goal-directedness. *How else* should we characterize the extensive process of storing and retrieving nuts across the seasons, or the many other clear striving behaviors (such as hiding, fleeing, reaching, leaping, nest-building, and so on) that the squirrel undertakes every day? It would be difficult to describe any of these behaviors except in terms that, at least implicitly, make reference to what purpose the behavior serves—what it is *for*.

If we scale down somewhat in complexity from the comparatively big-brained and complex-bodied squirrel, we can look at animals such as mosquitoes, whose bodies and behaviors, while simpler, are still astonishingly adapted to a particular lifestyle.

Figure 7.2: A female mosquito about to have a drink in order to nourish her eggs. Photo reprinted courtesy of James Gathany, Centers for Disease Control and Prevention.

Like most insects, a mosquito is born from an egg that its mother carefully laid in a relatively safe place in order to ensure that her hatchlings would flourish. In the mosquito's case, this is at the surface of a body of standing water. And, like most winged species of insect, when the larval

mosquito awakens into the world, its first actions are to alternately feed (in order to gather nutrients) and molt its exoskeleton (in order to make space for new growth) until it is large enough to undergo metamorphosis into an adult (in order to prepare for reproductive behaviors).

When fully formed, the adult mosquito climbs out of its pupal shell and up onto the surface tension of the water for the first time. It stands for a short while in order to let its soft body harden, crawls ashore, and then opens its wings in order to let them dry. Then the female mosquito, as we all painfully realize, spends the rest of its nights skulking about, seeking animals and humans whose blood she will extract with a syringe-like mouth. The male has more or less the same mouthparts, but it will use them to dine only on plant nectar in order to stay alive long enough to inseminate a female. So why is it only the female that needs to pilfer blood? Because the nutrient-rich fluid is used primarily to nourish its developing eggs in order that the whole story may begin again.

This description of the mosquito's lifecycle is intentionally riddled with "in order to" clauses in order to emphasize the creature's goal-directed nature, but those clauses can't simply be done away with. If we spoke of the series of activities that a mosquito goes through in its life without also explaining (or assuming) what each of those acts is for, we would be left with a lingering sense of curiosity, fueled by both the functional interdependency and the thermodynamic oddity of each of its behaviors: each part of these creatures clearly serves a role in their reproductive lifecycle, and each is used for (or is good for) accomplishing a certain task that contributes to the continuation of that lifecycle—each helps to organize the world in a certain manner, working against the statistical tendencies toward both energetic and material disorder. Everything seems to be done for a reason[268].

[268] I'm being rhetorically blunt here at the risk of offending anti-adaptationist sensibilities. Of course some parts and behaviors of creatures are non-functional, evolutionary accidents. The point here is not whether or not the functional parts and behaviors make up the lion's share; it is that, in terms of contributions to a continuing lifecycle, the functional parts and behaviors are the ones that matter, and the non-functional, non-adapted ones—however many there are—are largely irrelevant (unless they cost the organism too much).

In fact, every organism can be characterized in terms of this functional interconnectedness of its parts and behaviors, a state of affairs in which each stage of its lifecycle can be considered to do what it does *in order that* the next may be able to do what it does, and in order that, taken all together, the goals of survival and reproduction are achieved. And, in the case of the minuscule-brained mosquito—at least more so than in the case of the squirrel—this glaringly apparent goal-directedness cannot be psychologically attributed to what the organism was thinking or wanting. To emphasize this point, we can turn our attention now to a few examples where similarly discernible goal-directed behavior occurs but in creatures that lack brains altogether.

Photographs of sunflower fields are famously attractive with their rows and rows of large, bright canary-colored blooms. Such images are also remarkably easy to capture on any given sunny day. One might imagine that sunflowers, with vertical faces that inherently must favor one direction over any other, would be even more difficult to coordinate for a family portrait than people are ("Everybody look at the camera and say 'cheese'!"). To the contrary, though, the sunflowers line up in readiness for such photos all day long. They are virtuoso models that can transform any amateur into a directorial genius.

Of course, sunflowers are entirely unaware of and uninterested in the fact that they are being photographed. What actually accounts for the photographer's fortune is a common mechanism in many plants that botanists generally call phototropism (light-following) or sometimes, more specifically, heliotropism (sun-following).[269] The explanation for this behavior is straightforward: the plants that follow the sun do so in order to maximize the amount of solar energy they can collect and use for photosynthesis.

---

[269] More specifically, fully developed sunflower heads are *not* phototropic, but their leaves are, and when the plants' buds are young and green they follow the sun too. However, as the plant ages and the stem hardens and becomes woody while the flowers bloom, the mechanism that once moved the head back and forth becomes more limited and the heavy flowers tend to hang predominantly with their faces to the east . . . a result probably of the biasing fact that they spent all their youthful nights facing that direction, waiting for the sunrise. Still, that resultant bias can be attributed to the heliotropism of their adolescence.

Figure 7.3: A field of sunflowers facing to the east in the mid afternoon. (Image credit: An amateur photographer.)

As with the activities of squirrels and mosquitoes, those of plants are best described in terms of what purposes they serve. Heliotropism is not an outlier in this regard. Defense against insects and herbivores, the production of flowers (in order to attract pollinators), and the production of seeds, with a wide variety of mechanisms for endurance and dispersal, all serve various purposes in bettering the odds of producing the next generation. These things are done for a reason.

But of course we have little chance of chalking up these reasons to psychological intentions. Plants are classic non-psychological agents, according to most people's intuitions. They have neither brains nor ganglia, nor even a few distributed neurons. And while they have numerous molecular signaling systems[270], most people would agree that they have no hopes or desires, no beliefs about

[270] There is at least one entire journal devoted to the topic: *Plant Signaling and Behavior.*

possible futures, and thus no ability to make plans toward which they might strive. Plants might be *minimally cognitive*[271] in the sense that their behaviors may include simple, conditional, information-processed reactions to stimuli (if heliotropism is unconvincing, then think of the Venus's flytrap), but the goal-directedness we see in them is not akin to psychology. They are just too simple, too reflexive, for planning.

To climb down the complexity ladder further still, we can strip away the distractions of multicellularity and examine the behavior of single-celled organisms such as bacteria. When we zoom in for a close look we'll find that these organisms appear no less goal-directed than the plants and animals we've explored above[272]. The basic lifecycles of bacteria consist in feeding (in order to procure materials for the next steps), the synthesis of RNA, membrane lipids, and many other functional molecules (in order to repair damage and prepare for reproduction), and then the many ordered steps of binary fission undergone in order to produce two daughter cells that are then ready to begin the cycle once more. This is mechanically more modest than the lifecycles of multicellular creatures, but it is nonetheless vitalistic, and can equally be described as a series of functionally interconnected activities, each done *in order to* prepare for the next.

[271] The term "minimally cognitive" is borrowed from Randall Beer who uses it to refer to the simplest behaviors that we would characterize as being cognitively interesting. His first example is an artificial agent with a few simple sensors and motors (Beer 1996). In that and later work, Beer used an evolutionary algorithm to pursue minimal neural networks that could direct both simulated and real agents in performing simple discriminatory or behavioral activities (see, *e.g.*, Chiel, Beer and Gallagher 1999; Beer and Williams 2015).

[272] A testament to this resemblance is the fact that Anton van Leeuwenhoek, the microscopy pioneer and discoverer of microorganisms, first called them "animalcules", from the Latin for "little animals".

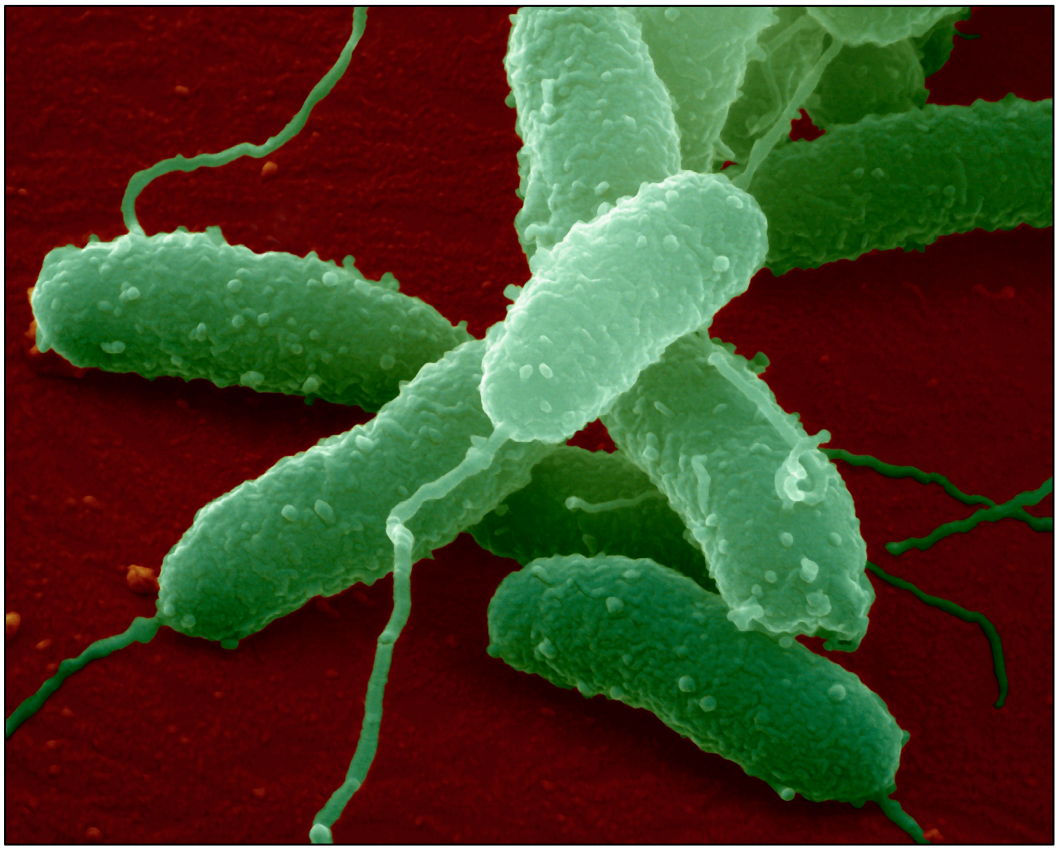


Figure 7.4: A colored scanning electron micrograph of the flagellated bacteria *Vibrio cholerae*. Zhu *et al.* (2002) discovered that *V. cholerae* use quorum sensing (described in the text below) to control gene-expression for toxicity virulence. (Photo taken by Juergen Berger, Max Planck Institute for Developmental Biology.)

In addition to the functional interconnectedness of their structures and behaviors, bacteria, much like animals, give the appearance of striving in many of their individual behaviors too. Howard Berg, one of the foremost contributors to our understanding of the biochemical mechanics of bacterial locomotion, notes that the mechanisms that contribute to their success in searching for food is best described as the *pursuit* of greener pastures. In his monograph "*E. coli* in motion", Berg (2004) describes the molecular mechanisms by which *Escherichia coli* use their propeller-like flagellum, driven by the world's tiniest mechanical motor, in order to swim upstream (to greener pastures) in a

concentration-gradient of nutrients, by using sensory information about their progress to direct them[273].

This ability to "sniff out" and dependably move towards nutrients, such as glucose, serine, or aspartate, and away from poisons, such as the chemical phenol, is called *chemotaxis* and it is shared in some form by many species of bacteria. In consisting of a sensorimotor feedback loop mediated by information processing, chemotaxis is a paradigmatic minimally cognitive behavior, and it is also an undeniably purposive-looking one. As Berg himself observes, "*E. coli* swims in a purposeful manner" (2004).

Another apparently goal-directed activity of bacteria is their social coordination through molecular communication. One such method that recently has been investigated in detail is called "quorum sensing". In this system, individual bacteria signal their presence to one another by releasing a special molecule that they can also detect. When the concentration of that molecular signal exceeds a threshold, the members of the colony recognize it as a quorum—a minimum number of votes that has been reached. This signifies that their numbers are great enough to effectively accomplish a mutually beneficial, but cooperative task, such as virulence, spore-production, or biofilm formation, activities that are much less effective or simply wasteful when

[273] The details are fascinating. Like many bacteria, *E. coli* has helical flagella (tails) that can act much like the propeller on a boat. As it turns out though, there is no rudder on these bacteria, and the environment they swim in is one of constant perturbation from Brownian motion (Brown 1828; Einstein 1905; Perrin 1909), making it a bit of a challenge for them to set their course in a straight line for very long, not to mention that they have no way to know which direction they are facing when standing still. The brilliant solution that evolution came up with to solve this challenge is a kind of hill-climbing search algorithm that exploits the simple information in a nutrient gradient. The flagella (which are attached somewhat variably between stern and midship) have two modes of operation. When they turn counterclockwise, they become more or less bundled together, cooperatively propelling the bacterium forward. When they turn clockwise, they disentangle from one another and work uncooperatively, causing the bacterium to tumble about unpredictably. Thus, with a simple switch in polarity in the motor, the bacteria can swap between two different kinds of activities that Berg calls "tumbles" and "runs". The trick is in the way these tumbles and runs are coordinated by information. The basic architecture of the algorithm is to simply alternate between tumbling and running, thereby making a little distance and then reorienting to a new unpredictable direction, but this alone amounts only to a random walk. The key to making *useful* progress is a one-bit sensory system in the bacteria, which turns these tumble-and-run cycles into something much like the game of "warmer/colder" that children play. The bacteria know only whether the current concentration of nutrient outside them is greater or lesser than it was a moment ago, and if it is greater, then a run is allowed to last longer, but if it is lesser, then a run will be cut short, and another tumble will take place. Thus the random walk is punctuated with longer runs in the right direction (and shorter ones when going the wrong way). The result is a fair bit of backsliding and sidestepping but, overall, the bacteria make forward progress (Berg 2004).

performed individually (Bassler and Losick 2006; Miller and Bassler 2001; Zhu *et al.* 2002). Cooperative social coordination of this sort would be difficult not to call goal-directed, especially when it has effects that serve what we see as being the purposes of the members of the population.

I'm going to review one more example now, because it is a borderline case that will help later in mulling over the relationship between being goal-directed, being animated, and being alive. The example is that of viruses, which most biologists consider neither to be organisms nor to be alive but which still, when we look at them closely, appear goal-directed, with lifecycles in many ways similar to those of organisms.

Every virus has a structure that is a variation on the following two-part plan: (1) a genome, consisting of either DNA or RNA, is packed into (2) a protective shell called a capsid, usually icosahedral (twenty-sided) or cylindrical in shape and made of numerous identical protein parts. Depending on the species of virus, this capsid may then be enveloped in a membrane similar to that of a cell, or it may have a tail (as many bacteriophages do), but some simpler viruses have only a capsid.

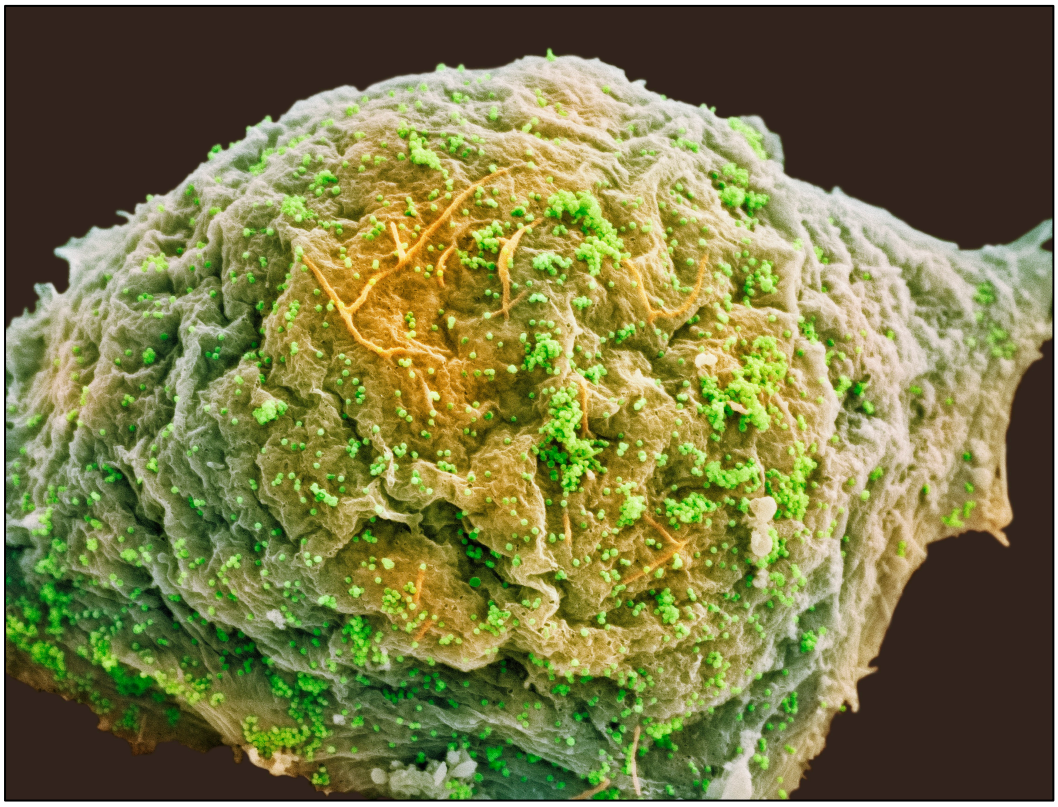


Figure 7.5: A colored scanning electron micrograph of human immunodeficiency virus (HIV) virions budding through a cell wall. The new virions are the numerous tiny green spots (Photo taken by Thomas Deerinck, National Center for Microscopy and Imaging Research, University of California, San Diego).

The reason that biologists typically refuse to call this type of structure alive is that, in the absence of a host cell, the virus is entirely inert. It does nothing. Unlike the much more complex structure of cells, viruses don't have their own metabolic pathways, nucleotide-replication mechanisms, or protein-synthesizing machinery, and so they can neither harvest energy nor even use any if they had it, in order to reproduce or further their own survival (by, say, repairing environmentally-sustained damage) [274].

[274] For an interesting borderline exception, see Häring *et al.* (2005).

In the context of a host cell, however, a *virion* (as individual virus particles are called) becomes another thing altogether. As soon as it comes into contact with the right kind of host, the virion springs into action, first attaching to the cell and then forcing its way in, to gain access to the tools inside[275]. Once it is inside, there are a number of strategies that it might take, depending on the species. For instance, viruses with latency strategies will inject their genetic material into the host's genome and then more or less hide away for a while, in order to allow themselves to be passively replicated and spread as a byproduct of normal cell division of the host. Other viruses will immediately set about building virion-production factories—new, localized cellular machinery of their own—which then produce new virions in a highly efficient manner. But most viruses will simply commandeer the cell's normal operating machinery in order to copy their own genetic code and synthesize their own proteins, which eventually assemble into new virions within the cytoplasm. The new virions are then borne out into the world beyond the cell either by rupturing the entire cell or, in the case of enveloped virions, through a final act of thievery, called *budding* (see Figure 7.5), in which they create their envelope by pushing through the cell's membrane and pinching off a bubble of it as they go[276].

Some scientists (*e.g.*, Bandea 1983; Forterre 2010) have suggested that we should see individual virions as *spores* or seeds, and only whole infected cells as viruses. What they mean is that, although the virion does not build the cell, once it commandeers it, then it effectively owns the machinery therein and it can be said that there is a new entity, comprised of the virion fused with its cellular home, for which these authors suggest we reserve the word "virus". Since these merged

[275] On review, it appears that many of a virus's behaviors would fit a description of minimal cognition. Despite being comprised of just a handful of molecules, the virion seems to use information to make decisions. For instance, the detection of the right kind of host cell by way of a sensor consisting of matching proteins, which then triggers a behavior—the entry mechanism the virion uses to get into the cell—seems strongly analogous to other sensorimotor feedback mechanisms in more complex organisms. And this is just one example of the virus knowing just when to engage in which activities—in an ordered lifecycle—in order to successfully effect its own replication.

[276] The most insidious thing about this act of stealing the door as they exit is that these virions then use that very same stolen door as their entryway to the next host cell by inducing the new host's membrane to fuse with it, thereby accepting the virion right into the factory.

entities are cellular and have features such as membrane containment, metabolism, and ribosomal protein-synthesis, Bandea and Forterre both consider them—but not their virions—to actually be alive by the same standards that other biologists use in describing cellular life. One obvious concern with this view is that the virion is not self-sufficient in that it does not contribute to the construction of the membrane of the cell that it comes to inhabit, nor much of the machinery that it comes to operate[277]. But at any rate, the proposal is an interesting one, not least because it is driven by a desire to recognize and make sense of the vitality or goal-directedness that these writers see in the many complex intracellular behaviors of virion-cell amalgams.

We've looked now at five examples that span a variety of the domains of life, from animals and plants to bacteria and even pseudo-living viruses. At this point, I think we can conclude that no matter where in the biological world we look, the behaviors of creatures that have lifecycles are going to give us the impression of being goal-directed. Partially this is because their vitality and their industrious projectivity resemble our own constructive, project-focused strivings; partially it is because each activity that organisms undergo seems as if it contributes to something; and partially it is because each activity within a lifecycle seems to make sense only in terms of the whole cycle—each seems as if it is done in order that the next may begin. Even when things turn out to have been done in vain, they still don't seem to have been done for nothing.

Well before people understood cells or had any other productive biological theories, we labeled organisms as alive and grouped them together as biological objects simply because *they displayed vitality*. We knew that bacteria and molds had something in common with sunflowers and mosquitoes and squirrels and humans because they all appear to have their own interests that their behaviors seem to serve; they all behave in a way that appears to us to be vitalistic, agential, and

[277] Although in this case too, there are some partial exceptions in which viral genetic material codes for the construction of a set of operating mechanisms for virion construction, or even for some photosynthetic components that replace those of the cell and produce the energy necessary for virion construction (see *e.g.* Bragg and Chisholm 2008; Miller and Krijnse-Locker 2008; Novoa *et al.* 2005).

goal-directed. We see the same thing whether we are looking at large multicellular organisms, the microscopic organisms that Leeuwenhoek first called "animalcules", or even the internal workings of an individual cell. It is an observation that is plain to any of us, from the prescientific philosophers of thousands of years ago[278] to two-year-old children today. We see vitality and we see purpose.

[278] Recall the quote from Aristotle, cited earlier: "It is absurd to suppose that purpose is not present because we do not observe the agent deliberating." (*Physics* II.8)

B. The Teleologist's Dilemma

Minds are not bits of clockwork, they are just bits of not-clockwork. As thus represented, minds are not merely ghosts harnessed to machines, they are themselves just spectral machines. . . . Now the dogma of the Ghost in the Machine does just this. It maintains that there exist both bodies and minds; that there occur physical processes and mental processes; that there are mechanical causes of corporeal movements and mental causes of corporeal movements. I shall argue that these and other analogous conjunctions are absurd.

—Gilbert Ryle (1949)

When what a person does is attributed to what is going on inside him, investigation is brought to an end. Why explain the explanation? For twenty five hundred years people have been preoccupied with feelings and mental life, but only recently has any interest been shown in a more precise analysis of the role of the environment. Ignorance of that role led in the first place to mental fictions, and it has been perpetuated by the explanatory practices to which they gave rise.

—Burrhus Frederic Skinner (1974)

Beginning just over a century ago, a generation or two of psychologists found themselves cornered by a dilemma: The alternatives were to believe in a kind of extra-physical mind-stuff—the view called dualism—or to face what seemed like the perverse task of explaining the seemingly immaterial, subjective mind in objective, material terms. The behaviorist's resolution at the time, which today is unsatisfactory to most, was to avoid landing on either horn of the dilemma by claiming that the mind simply does not exist—that it is an illusion. Without something to explain, the concerns of both alternatives dissolved. *Behavior* could still be explained by material "laws" such

as classical and operant conditioning, reinforcement learning, and Thorndike's "law of effect", but such explanations eschewed any reference to *mental states, beliefs, desires,* or *minds*. Writing off these latter concepts as fictions seemed to simplify the psychologists' theoretical work, and the ensuing success of their theories in describing behavior gave them great confidence.

The modern teleologist faces much the same kind of dilemma. On the one horn, vitalism, which posits a kind of non-physical life-stuff or goal-stuff, is to biology and teleology what dualism was to psychology. And on the other horn, it seems equally perverse to attempt to account for the seemingly immaterial and, again, *subjective* notions of goal-directedness and normativity in objective material terms. How could mere mechanistic physics and biochemistry account for strivings, for reasons, and for plans, not to mention a set of norms by which to evaluate the achievement of those strivings and plans?

Not surprisingly, without easy answers to these kinds of questions, the modern philosopher and scientist usually escapes the teleologist's dilemma in the same way the behaviorists escaped theirs. They deny or ignore the existence of non-psychological goals in the world, and give theories of much of what they observe in terms of a related concept (*function*) which itself seems to them explainable in goal-free, objective, material terms—terms such as causal roles, dispositions, etiology (selected effects), and natural and human design.

André Ariew takes himself to speak for the field in endorsing this kind of response to the teleologist's dilemma:

> "Functional explanation" is our chosen term because "teleological explanation" is thought to imply backwards causation or bizarre ontological categories (for example vital forces) attributable to the teleological theories of Plato and Aristotle.

> Functional explanation is not so imbued and hence, as opposed to teleology, is an appropriate topic for naturalistic analysis. (Ariew 2002)

The idea that any topic could be deemed "inappropriate for naturalistic analysis" is a strangely unscientific one. All patterns in the world that we (or any other agent) can in principle perceive are ripe subjects for analysis. By their very nature, as patterns, they are immediately available for making and testing predictions. On one hand, perhaps the pattern will turn out to be an illusion, but if so, analysis can only help to expose the source of that illusion. On the other hand, perhaps what we perceive is a real pattern, out there in the world, and we simply have yet to successfully wrap a theory around it (Dennett 1991). In either case we will come to understand the pattern more deeply through naturalistic analysis.

Now behaviorism no doubt provided us with some abidingly interesting and valuable notions, but it ultimately failed to explain fully the subjects of psychology and behavior. By turning its back on some of the fundamental patterns of its field of inquiry—thoughts, beliefs and desires—not only did it leave such patterns themselves completely unexplained, but furthermore it was unable to account even for human behavior as successfully as an alternative body of theory (cognitive psychology) that does accept the roles played by thoughts and beliefs in the determination of behavior.

Likewise, the recent decades of inquiry into function and design have also provided biology and philosophy with many useful notions. However, in largely ignoring goals, all this work leaves still unexplained not only the observed goal-directed character of organisms—their striving, their perseverance, their directedness, their vitality, and their normative nature—but also, I will soon contend, it leaves unexplained the way in which functioning deeply depends on that goal-directedness and, thus, in which it cannot be fully explained without reference to goals.

Now, in laying out this analogy between behaviorist psychology and functionalist teleology, I'm not suggesting that *any* argument based in ontological doubt is systematically wrong. Some patterns in the world do in fact turn out to be illusions, and our doubt is what leads us to eventually determine that about them. What I am trying to suggest is that the reason that such doubt exists, both in the case of minds for behaviorists and in the case of goal-directedness for biophilosophers, is equally poorly founded. In both cases, the doubt is not based in good evidence that there may be an illusion—there is no investigable pattern of exceptions or blatant context-relativity that seems to characterize the doubted pattern. Rather, these patterns are doubted to exist only because *explanation of them proves difficult*. The theoretical strategy employed is to assume that if the pattern can't easily be made to fit with our rather well-tested scientific worldview, then it must not exist.

That strategy might seem to make sense since it appears to work wonderfully for such patterns as goblins, ghosts, and gods; but the comparison is not fair. In cases of that kind, the bulk of the prior observational evidence is against the existence of such phenomena. The patterns are insubstantial, fleeting, ethereal . . . not *obvious*, and hence the burden rests upon the believer's shoulders to prove to us why we should believe. As the Bayesian scholar would put it: evidence *for* their existence must be overwhelming, even miraculous, in order to overcome the prior evidence that they do not exist.

For cases like beliefs and goals, though, the bulk of the evidence clearly supports their existence. Nothing could be more obvious than the fact that we have minds and thoughts and beliefs that direct our behavior. It is clear as day to all of us today, and it could only have been clear as day to the thinkers of the early 1900s as well. Our minds are there, "staring us in the face", every moment of every day. Coming to discover such patterns did not require mental gymnastics and explanatory toil, but it was rather the disguising of these patterns, in order to eradicate them, that required years of (ultimately unsuccessful) effort by behaviorist thinkers.

Similarly, as Rosenberg said of organismic aims and purposes, "Nothing could be more obvious" (1985:43)[279]. *Nothing could be more obvious.* When we look to the natural world, the biological world, there are goal-directed behaviors being performed everywhere. The squirrel, the sunflower, the mosquito, the bacterium, and the virus are just a handful of examples, but they aren't special cases. Every organism on the planet has behaviors that are goal-directed and this brute fact—this pattern of regularity—can be clearly observed by every two-year-old child discovering the liveliness of insects, birds, animals, and plants in their world and seeing something in their animacy that clearly distinguishes them from rocks and puddles and so on (see also Bertenthal 1993; and Rochat *et al.* 1997; as well as Piaget 1929).

Like beliefs and minds, goal-directedness is also staring us in the face all the time. And since the bulk of the evidence clearly supports the existence of these patterns, then contrary to how we treat goblins, ghosts and gods, it rests squarely upon the *non-believers'* shoulders to prove to the rest of us why we should disbelieve in the goal-directedness we see. If one intends to wipe out a clearly observed pattern from their ontology for the sake of theoretical simplicity, then one had better have a good reason to believe that the pattern is indeed an illusion. That it is merely difficult to explain will not suffice.

[279] Though, strikingly, Rosenberg himself is a selected-effects theorist and turns out to be a goal eliminativist.

C. Goal Eliminativism

Teleology is like a mistress to a biologist: he cannot live without her but he is unwilling to be seen with her in public.[280]

— J.B.S. Haldane

The term "goal eliminativism" (alternatively: "teleological eliminativism") can be used to refer to the common belief that any goal-directed behavior we observe in the world is only *apparent*. This differs importantly from earlier uses of the word "eliminativism" in the functions literature, wherein it was used for instance to classify views that "reject the intuition that a real difference exists between functions and dispositions" (Enç and Adams 1992). Put simply, function eliminativists would have us disbelieve in functions while goal eliminativists would have us disbelieve in goals[281].

The theory of goals underlying goal eliminativism is typically not that there are none. We could call that thesis *out-and-out* goal eliminativism and perhaps it was held by some of the behaviorists who refused to believe in mental states including psychological goals, but it is rarer nowadays. Instead, most modern versions of goal eliminativism are roughly characterized by the belief that the only goals that truly exist are psychological goals, and that it is the non-psychological remainder of observed goal-directedness that is illusory. There are a number of variations on this stance, but for all of them the behaviors of the sunflower and the bacterium are not attributable to any intrinsic goal-directedness of those organisms (while those of *humans* are). Commonly it is taken that the behavior of these other organisms is *apparently* goal-directed, and that this appearance can be

[280] I don't mean to endorse this glaringly sexist comment. Still, I felt compelled to include it here because the underlying theme (of even a gender-neutral translation of it) is exceptionally illustrative of the old but long-running and I think particularly problematic eliminativist stance that biologists take.

[281] Some might be tempted to classify *me* as a function eliminativist. If they did so, I would find it difficult to argue otherwise but I would like to emphasize that, while I hope scientists eradicate the notion of a "proper function" from our *scientific* vocabulary, I find the *common* notion of an item having a function to be as exceptionally useful as the notion of an object having a weight or a color.

explained by some superficial analogy between their behaviors and those of psychological agents such as us. We could call this kind of view *partial* goal eliminativism (as opposed to the out-and-out version), since its adherents do believe in some but not all goals; I will call it just plain eliminativism, as it is the only form that we really run into nowadays[282].

Let's look next at a variety of eliminativist claims in order to clearly characterize the view and the ways it has come to be dispersed. Today, eliminativism is the default position held by the vast majority of scientists and philosophers.

[282] Allen and Bekoff (1995) have already branded this mentalistic kind of partial eliminativism "teleomentalism", in order to contrast it with what they call "teleonaturalism", the latter being the class of theories that claim to give not a psychological, but a natural source for purpose or function. There are, however, some eliminativist (and thus teleomentalist) views, both brazen and closeted, amongst the views that Allen and Bekoff place into each of their categories. For instance, Davies' (2001) causal-role theory of function (another defense of Cummins' view) is archetypically teleonaturalist but at the same time he offers an avowedly teleomentalist explanation for goals. Allen and Bekoff's taxonomy works well to classify the *functions* debate that we looked at earlier; but I'll stick to the word "eliminativist" to describe teleomental views on *goals*.

D. Eliminativism Everywhere

Nearly all biologists and philosophers agree that biology is fundamentally non-teleological.

—Michael Ruse (1971)

David Hanke, a botanist at the University of Cambridge, is one recent writer I've run across who is absolutely captivated by the teleologist's dilemma. He finds neither horn acceptable and so argues forcefully for eliminativism:

> Biology is sick. Fundamentally unscientific modes of thought are increasingly accepted, and dominate the way the subject is explained to the next generation. The heart of the problem is that we persist in making (literally) sense of a world that we now know to be senseless by attributing subjective values to the objects in it, values that have no basis in reality. (Hanke 2004)

If one reads Hanke (2004) further one finds that the "subjective values" he is referring to are the norms of purposiveness that underlie the way we see biological behavior as functional and goal-directed. He worries that we are seeing reasons, in biology, where none exist. More precisely, he thinks that there are objective "how come?" reasons for science to explore and discover, but that there are no subjective "what for?" reasons in our world, and so to speak in such purposive terms is fallacious. Continuing, he writes, "The purpose of any object is entirely subjective because *purpose has no real existence outside the mind of the animal thinking of it*" (2004, emphasis added). So, for Hanke, there can only be psychological goals and purposes in our world. In the next few sections of this chapter, I think we'll see that Hanke is not nearly as alone in these beliefs as he thinks he is.

André Ariew appears to be more circumspect, but also ultimately prefers an eliminativist stance of one sort or another. In his exposition of classical Greek teleology he concludes "*if* biology has an ineliminable teleology, this is not so bad as long as it is one of the more restrained Aristotelian versions of teleology" (2002, emphasis added). And he clarifies his idea of "the more restrained Aristotelian versions" with an assertion that I reviewed earlier: "teleology pertaining to natural organisms is . . . non-purposive (though seemingly so)" (2002). Reading further, we find that Ariew would accept an "Aristotelian explanation for the existence of traits in terms of their usefulness"—that is, a Darwinian theory of function, but not a theory of goal-directedness (Ariew 2007). As with Hanke, there is no *real* purposiveness in any organisms other than psychological human agents—Ariew would have us accept either a goal-free materialism or a weak version of Aristotelian teleology in which purposiveness is only "seemingly so".

Employing the adverbs "seemingly" or "apparently" to preface "goal-directed" and "purposive" is one of the clearest signs of an eliminativist stance. It is also a nearly universal practice today. Ariew (2002) notes that qualifiers of this kind have been used at least since Aristotle distinguished his four causes. Indeed the trend is so widespread that the term "teleonomic", introduced by Pittendrigh (1958), has come to be embraced and used widely as synonymous with "seemingly" or "apparently" goal-directed (though Pittendrigh himself hadn't intended it this way)[283]. The long list of quotes that follow showcases the diversity of disciplines across which purpose is seen to be "seeming" or "apparent".

Dan McShea has recently offered a new theory of teleonomy in terms of hierarchical structure. The view is eliminativist, though.

[283] He simply meant to distinguish a deterministic kind of teleology from cosmic or theologically connoted teleology or backwards causation. Pittendrigh was not an eliminativist; he did in fact believe that biological objects are goal-directed, and was somewhat convinced by the cybernetic theory of goals. (See the endnotes of Mayr, 1974/1988, for correspondence between Mayr and Pittendrigh detailing this issue).

> How shall we understand *apparently teleological systems*? . . . Here I argue that all *seemingly goal-directed systems—e.g.*, a food-seeking organism, human-made devices like thermostats and torpedoes, biological development, human goal seeking, and the evolutionary process itself—share a common organization. (McShea 2012a, emphasis added)

Carl Craver, another philosopher of biology also recently offered his eliminativist view:

> I claim that *the causal structure of the world is disenchanted and purposeless*. Mechanistic and functional descriptions, in contrast, presuppose a vantage point on the causal structure of the world, a stance taken by intentional creatures when they single out certain preferred behaviors as worthy of explanation. Specifically, talk of functions and final causes is not legitimized by or reduced to privileged kinds of etiological histories (though some functions have such histories) or to certain special effects of the item in question. Rather, they are imposed from without by creatures seeking to understand how a given phenomenon of interest is situated in the causal structure of the world. (Craver 2013, emphasis added)

McShea and Craver may be some of the most recent theorists of teleology, but they also exemplify an old trend. In introducing their anthology of essays on "Nature's Purposes", Allen, Bekoff, and Lauder hedge their bets too.

> Biology is unique among the natural sciences in licensing *apparently teleological statements about design, purpose, and adaptive function*. Teleological thinking originated from two

> views, both of which are assumed to have been discredited in physics, chemistry, and the other natural sciences. (Allen, Bekoff, and Lauder 1998, emphasis added)

Laubichler, also writing about teleology in terms of function, says:

> This ontological distinction between mechanistic processes and "additional laws of nature" also sets the stage for the binary opposition between "reductionism" and "holism" that has often crippled any constructive dialogue about the interpretation of *seemingly goal-directed or purposeful biological phenomena*. (Laubichler 1999, emphasis added)

In approaching the teleologist's dilemma, Sommerhoff attempts to navigate between the horns: "We see therefore that a definite answer can be given and that it is neither the teleological answer of the vitalists nor the skeptical answer of the mechanists." (Sommerhoff 1969). Ultimately, this turns out to be an eliminativist view.

> The most distinctive characteristic of the behaviour of higher organisms is its goal-directedness, *its apparent purposiveness*. In fact, it is largely through this *apparently teleological nature* of their activities that living organisms betray their exceptional organization. (Sommerhoff 1969, emphasis added)

Also in another passage:

> If we abandon scientific exactitude and provisionally attempt to express these fundamental characteristics of living systems in non-scientific and largely metaphorical language, we may say that they consist in the *apparent purposiveness* of vital activities and in the manner in which this *apparent end-serving or goal-seeking quality* integrates the part events of living systems into the self-regulating, self-maintaining, and self-reproducing organic wholes which we recognize as living individuals. (Sommerhoff 1950)

Given these quotes, one might be surprised to find that Sommerhoff does offer a theory of goal-directedness. But he is not inconsistent. His theory is given in the tradition of the cybernetic theory, and, like other authors in that stream, he is only prepared to say that the goal-directedness he is describing is illusory, the result of an often complicated, but ultimately cybernetic, mechanism that amounts to little more than the objective mechanisms in things such as thermostats and homing torpedoes. It is not truly normative, subjective, or value-laden.

The eliminativist trend is not limited to philosophers of teleology. Other writers in biophilosophy display the same commitment.

> Memes, like bacteria, create their own *apparent goal-directedness* because they form a selective system with replicators whose permanence transcends individual bodies. (Frank 1996, emphasis added)

> The most salient feature of organisms is adaptation, the *seeming goal-directedness* that makes organisms different from merely physical entities. (Queller and Strassman 2009, emphasis added)

Medicine is a biological subfield and so it is subject to a somewhat similar culture of beliefs. For example, in describing the cause of the coughing reflex caused by stroking the trachea, Goldstein writes,

> The *apparent goal-directedness* and coordination of this reflex does not imply either that the patient senses the suctioning catheter or coughs voluntarily, just as the *apparent purposiveness* of a thermostat does not imply that the thermostat is either conscious of or "wants" to maintain the temperature of a room. (Goldstein 2006, emphasis added)

Outside of biology and philosophy, we still find many scientists making similar claims. Though, in some of these contexts, the ontological doubt may seem somewhat more justified. For instance, from roboticists, we get the following quotes:

> We experimentally examined whether differences in the manner of interacting with a moving robot (operating it or only observing its movements) influenced one's perception of the robot's animacy and, if so, whether the strength of this influence depended on the *apparent goal-directedness* of the robot's movements. We found that people only observing the robot perceived it most animated when its movements seemed most goal-directed but that people controlling the robot perceived it more animated when $1/f$ noise made its movements seem less goal-directed. (Fukuda and Ueda 2010, emphasis added)

> A key thing to note with these robots is the ways in which *seemingly goal-directed* behavior emerges from the interactions of simpler non goal-directed behaviors. (Brooks 1990, emphasis added)

> The implemented architecture is used to control a simulated robot, and a classic experimental paradigm in which rats performed *apparently goal-directed* action selection is emulated. (Shanahan 2005, emphasis added)

And from evolutionary psychology, we have:

> Motivated responses may occur without the activation of any cognitive representation (conscious or nonconscious) of goal constructs. *Seemingly goal-directed* cognitions may simply be triggered automatically in response to the perceptual recognition of certain situations. Again, the suggestion is that this automatic triggering process occurs as a result of the evolutionary history of the species. (Schaller 2003, emphasis added)

Altogether, there appears to be a rather widespread tension among the scientists cited here: The pattern of goal-directedness must be as obvious to all of them as it is to Rosenberg—as obvious as it has been to any of us since the age of two. It is a pattern that they all see, that they find themselves compelled to talk about in scientific discourse and yet, due to a fear of being impaled

upon one of the two horns of the teleologist's dilemma, they cannot refrain from either equivocation or blatant disbelief[284].

[284] There are a handful of notable exceptions, including the scientists whom I've quoted in the epigraphs at the start of this chapter as well as Walsh (2009), who argues at length that biology cannot do without the explanatory power of goal-directedness; Deacon (2013), whose ideas prefigure my own in some ways; and Nagel (2012), who argues that natural teleology is the only acceptable answer to explain consciousness, cognition and value. Nagel stops short of giving a theory—he just suggests that one is necessary—but the argument is certainly iconoclastic, given the atmosphere we have been reviewing in this chapter. I also find his argument that teleology underlies value to be mistaken—as we'll see in the coming chapters, I prefer to frame it in the reverse manner.

E. Teleomental Eliminativism

If we say that we "try" to catch a fly, we regard this as a perfectly legitimate use of the verb "to try". But if we next say that the fly in the hollow of our hand will "try" to escape, our modern scientific training intervenes and warns us that in the second case we are committing an illicit anthropomorphism. The fly, we are warned, is not a conscious rational agent, and therefore does not in any literal sense "try" anything. This rigour of thought is very laudable. Yet in spite of these wise injunctions the incontrovertible fact remains that there is a unique something about the observed behaviour of the fly which quite emphatically invites this anthropomorphism and which renders this behaviour far better suited to such an analogy and teleological conception than, say, the behaviour of a falling stone.

—Gerd Sommerhoff (1950)

The eliminativist view that psychological goals are the only real goals is implicit in most of the uses of "seemingly" or "apparently" to describe goal-directed behavior, but it is also offered explicitly, and in detail, by the theorists that Allen and Bekoff (1995) call "teleomentalists" (see footnote 282, p. 388). Theorists of this persuasion illustrate the ways they think that psychological goal-directedness can account for the perception of non-psychological goal-directedness.

Ducasse (1925), for instance, claims "only the acts of entities capable of beliefs and desires, are capable of being purposive" and he requires further that "causation by that belief and that desire jointly" is essential in order to speak of an act as being purposive. At the time of his writing, it was uncommon to think of animals as possessing beliefs and desires, and Ducasse appears to have agreed with this verdict: he suggests "when a squirrel stores away food . . . it is *not a purposive act*" (1925, emphasis original)[285]. He defends his view by appealing to the threat of backwards causation:

[285] Though, with respect to human beliefs, Ducasse apparently diverged from the behaviorist school of thought, popular at the time.

> How, indeed, could a fact that has not yet occurred explain, *i.e.*, be a possible cause of, a fact that has already occurred? And it is here that the teleological temptation comes in: Obviously, whispers the Devil, only if an intelligence aware of the contingency of the second upon the first, and desiring the occurrence of the second, is thereby moved to bring about the first! (ibid, p. 153)

Ducasse undoubtedly would have felt the same way about our sunflower, bacterium, and mosquito examples as he does about the squirrel. For him, non-psychological goals only have an illusory appearance of being goal-like.

Woodfield's form of teleomentalism holds quite similarly "the [goal-directed systems] that do not have minds are included [as being goal-directed] because they are similar to the ones that do [have minds]" (1976:163). Woodfield notices the similarity between biological and psychological goal-directed activity, but he means to downplay that similarity in order to consider the mindless systems to not *really* be goal-directed. His strategy would be workable if we knew just how and why the mind-bearing systems came to be goal-directed and if we knew that that same structure—whatever it is—were not a part of the mindless systems. We could then conclude that the similarity is only superficial. But the study of cognition and affect is a fledgling field, only now beginning to explore how human goals and decision-making arise from the mind, brain, and body, and the picture is still out of focus. Given our relative ignorance, it might be more productive to use the opposite strategy from Woodfield's: to use the similarity between mindless and mind-bearing goal-directed systems to guide our inquiry about just what structures might be shared between them, rather than assuming those similarities are unimportant.

Lowell Nissen (1993, 1997) gives the most recent teleomentalist view of this sort. For him there is goal-directedness only if an intentional agent intends the goal. Nissen claims that when

people invoke goal-directedness to explain the behaviors of organisms, they are imagining natural design to require a "programmer" (he seems to borrow Mayr's term) who intends the organisms to operate as they do and whose goals we are actually talking about. Since Nissen rejects the idea of a creator, he then concludes that explanations that appeal to goal-directedness in this way amount only to false claims; there is no goal-directedness in non-human organisms.

In contrast to Nissen, there are those teleomentalists who do believe that the functions and goals of organisms are quite literally derived from the goals of a supreme psychological agent—a creator that designed and built (or programmed) the world to work just so (*e.g.* Plantinga 1993). We can call these people creationists or, following Allen and Bekoff (1995), "literal teleomentalists". The creationist claim of course will depend crucially on an extensive and sustainable defense of the existence of the imagined creator[286].

Some thinkers, of course, take the view that natural design is sufficient to metaphorically play the role of a designer (*e.g.* Dawkins 1986; Kitcher 1993) and that this is good enough to account for designedness and thus function. Allen and Bekoff call this strand of thought "*metaphorical* teleomentalism" since it doesn't explicitly require any intentional or mental aspect. In chapters IV and V, however, I argued extensively against taking design as a basis for function, and there is no need to repeat those arguments now.

The primary concern I have with any real teleomentalist views (the literal creationist version and the version advocated by Nissen, Woodfield and Ducasse, but not Dawkins' metaphorical variety) is that they all rest upon notions of intentionality (including psychological goals) without defining those terms. That is, they require a goal-directed agent from the start (either a human or a god) without explaining how that agent came to be goal-directed. Such a view leaves the original

[286] Since it will be a long while before we'll see such a defense, I think the reasonable course of action here is to continue examining other theories of goal-directedness. And, if, in the meantime, we find a complete natural theory of teleology that makes the creationist view superfluous, then we should provisionally accept it until we one day have evidence of a creator.

problem of *what goal-directedness actually is* simply unexplained. The question is not answered, but rephrased: What is it about minds that makes *them* goal-directed?

Of course, we have plenty of evidence to accept the idea that minds issue forth *somehow* from the complexity of (embodied) brains, and to presume that perhaps the explicit cognitive goal-directedness of humans can be explained, *somehow*, as a result of our emotional system combined with reasoning and "executive function" and perhaps culture and experience. But it remains to be seen just what all of those things really consist of and just how they might produce goal-directedness. I find it substantially compelling to think that the cognitive capacities that make psychological agents goal-directed actually result from those agents being biological organisms in the first place. After all, our central example of a psychological agent—ourselves—is, if anything, a biological organism. This is why I am rather more convinced by a logic reversed from the teleomentalist's—one that accounts first for biological goal-directedness in some kind of natural terms, and only then for how the psychological goal-directedness of humans is a special case of that more general and fundamental answer.

F. Function Theorist Eliminativism

According to neo-teleologists, as I shall call them, we have *hearts because of what hearts are* for.

—Robert Cummins (2002)

There really aren't any purposes in nature and no purposive processes either. It's just one vast network of linked causal chains. The notion that Darwinian natural selection naturalized purposes is just a way of sugar coating its bitter pill.

—Alexander Rosenberg (2013)

Many of the theories of functions that we looked at in the previous chapter are also either openly eliminativist or closet eliminativist. It will be worth reviewing them to see how. Perhaps the most brazen eliminativists amongst philosophers of function are the causal-role theorists. While Davies accepts functions and rejects the label of "function eliminativist" (that he anticipates Enç and Adams, 1992, might apply to him), he is nonetheless a *goal* eliminativist:

> I thus reject the premise that we must make room for any such teleology [—genuine purposes or norms of performance—] in our theories of natural traits. (Davies 2001, p. 49)

> [Goal-directedness] is important insofar as it accounts for our temptation to see some objects as more functional than others; but it is not among the conditions necessary for the attribution of a . . . function. (ibid, p. 155)

On Davies' view, functions in natural traits don't require the existence of purposes or goals or reasons in the world. He offers a teleomentalist thesis that our perception of goal-directedness is an illusion:

> I want to suggest . . . that certain psychological capacities and limitations incline us to see some objects as more functional than others. . . . [I aim] to suggest by way of example how to diminish the intuition we have that much of the biological realm is purposive. (ibid, p. 75)

Davies doesn't offer any convincing argument of how (or why) our psychological capacities might actually bring about the illusion of presenting false purposiveness to us—he cautiously and explicitly defers that task to psychologists—but later in this chapter I will perform a version of the exercise for him. After presenting a likely mechanism for such an illusion, I will suggest that we have good reason to believe that its scope of operation is limited and that instances when it does operate are typically detectable, so that we need not take the radically skeptical view that the illusion might be pervasive.

Cummins also refuses to allow purpose or goals to play any part in his functional analysis view. He likens the SE analysis, which he calls "neo-teleology", to teleological mechanics—the supposed purposiveness in phenomena such as falling stones and planetary orbits—and then argues that, like the animistic purpose that was once seen in these mechanics, the purpose seen in biological traits can be considered explanatorily irrelevant if his causal roles are sufficient to explain functions.

> Biological traits once explained by a teleology grounded in appeals to the intentions, plans, and actions of a creator have, in discerning minds, given way to appeals to

> evolution generally, and to natural selection in particular. Neo-teleologists want to read this as the discovery of a legitimate grounding process for a teleological explanation of these traits. I am inclined to read the same intellectual development as analogous to what happened in mechanics and developmental biology: not a vindication but a replacement. (Cummins 2002)

While I agree with Cummins that grounding (basing our theory of) teleology in natural selection fails as does grounding it in creationism, I think he may be too quick to toll the death knell. Teleology was properly wiped out of mechanics because scientists discovered laws by which the behaviors of physical phenomena could be described without any evaluative norms, but the same is not true of biological, psychological or artifactual phenomena—we simply cannot avoid at least sometimes speaking of organisms in terms of what is good or bad for them. Of course, Cummins claims that his functional analysis provides a way to look at functional phenomena without invoking any concepts of goal-directedness or value but, as I argued in Chapter V, the job only seems to be accomplished through a bit of sleight of hand on his part. The card is still in the deck; he's only made sure it's no longer on top.

Causal-role theorists wear their eliminativism on their sleeves, as do some of the other function theorists cited earlier (*e.g.*, Allen, Bekoff and Lauder 1998; Ariew 2002; Craver 2013; Laubichler 1999; McShea 2012), but some form or another of perhaps less blatant eliminativism is a wide trend amongst other philosophers of function too. Probably the most obvious sign of this is the contrast between the immense scale of the literature on function and the relatively little modern discussion of goal-directedness. Additionally, coming from just about all corners, there seems to be either animosity or indifference towards the only modern goal-based theory of function, even

though this theory (Boorse's GC analysis) can, taken loosely, be independent of any particular theory of goals.

Selected-effects theorists seem for the most part to be teleologically agnostic, if not eliminativist. Most of their theories are given in terms that simply make no mention of purpose or goal-directedness. Many of them make no mention even of normativity (*e.g.*, Wright 1973; but see also Wright 1976). As I noted in the previous chapter, Millikan's SE functions are comparatively but not evaluatively normative so she might see herself as giving a kind of teleological theory of function[287], but still, the comparative, historical norms of selection only allow us to say that a trait is or is not doing what its ancestors had done. They don't answer inquiries that are interested in "for the sake of" or "in order to" or about what purpose something serves.

Perhaps the most plainly eliminativist SE theorist is Godfrey-Smith, who argues that Wright's (1973) introduction of the function–accident distinction "disposes of the whole range of analyses of functions based upon contributions to goals" (1993:197). Godfrey-Smith continues: "a mere contribution to a goal is not a function unless it is not fortuitous, unless this contribution explains why the thing is there. But this requirement of explanatory salience is apparently now bearing the whole weight of the concept of function, and *goals drop out of the picture*" (*ibid*, emphasis added). Of course, on Godfrey-Smith's argument, goals drop out of the picture only if one takes the notions of "proper functions" and the function–accident distinction seriously but, as we've seen, these are ideas that I seriously doubt.

As we saw in the previous chapter, some of the theorists whose views fall under the "survival and reproduction" (SR) branch of the RD analysis are also eliminativist. Canfield says, "If, then, the word 'useful' which appears in [my preliminary sketch] can be replaced by words which are clearly non-teleological, there will be strong reason to believe that the teleological notions occurring

[287] Indeed, Millikan claims, "The things that have 'proper functions' do seem to coincide with things (omitting God) that have, in ordinary parlance, 'purposes'." (1984).

in functional analyses can be dispensed with." This ambition to find a schema by which one could "translate away" any teleological terms, by trading in terms such as "in order to" for replacements such as "and thereby", was widespread at the time (Canfield 1964; Nagel 1961; Pittendrigh 1958; Ruse 1971), but has been convincingly argued against by a few authors who show that translations of this sort inevitably lose an important part of the original meaning, particularly with regard to the question of what an event or item is *for* (Ayala 1970; Beckner 1969; Mayr 1974). Mayr gives the following example.

> The Wood Thrush migrates in the fall into warmer countries *in order to* escape the inclemency of the weather and the food shortages of the northern climates. (Mayr 1974)

If we were to replace "in order to" with "and thereby", Mayr says, "we leave the important question unanswered as to *why* the Wood Thrush migrates" (1974). Beckner gives a series of examples in which the phrase "in order to" does not appear but is still implicit. When we understand these claims, we *tacitly* bring in our knowledge of goal-directed behavior and make the (correct) assumption that the subjects of examples T1–T4, along with other agents implicit in T5 and T6, are behaving in an agentive, goal-directed manner.

(T1) "Vultures break open eggs with stones."

(T2) "Myrtle warblers migrate in the spring into regions of abundant food."

(T3) "The missile swerved toward its target."

(T4) "He acted out of avarice."

(T5) "The arms of the Dean's chair are upholstered."

(T6) "The sight of a barracuda releases an escape reaction in anchovies."

(This list adapted from Beckner, 1969)

Despite lacking an "in order to" clause these are all teleological claims that compel the reader to think in terms of goal-directedness. We immediately wonder or assume what the vultures break open the eggs *for* and *why* the myrtle warblers migrate and so on. Beckner's (and Mayr's) point is that we should not be fooled by these translational-theorists who are avowedly eliminativist but whose lexical chicanery only eliminates certain *words*, not the *concepts* behind them.

Even the maverick theorists who defended the three unpopular theories are not all unequivocally willing to defend goals. Boorse is, and Mayr comes close[288] but Bedau, who supports the VE theory, presented his three grades of teleology in order to separate what he sees as "real" (for him, psychologically-based) teleology from what he takes to be biological pseudo-teleology. He says, "Grade three teleology *seems to be* present in biology, but only its grade two close cousin *really exists*" (Bedau 1992b, p. 284, emphasis added).

Altogether we find few amongst biologists, psychologists, roboticists, philosophers, and even the brand of function-theorizing philosophers that we could almost call "teleologists" who are willing to accept that there are real evaluative norms or real goal-directedness in our world. Today, thinking of purpose as being merely "apparent" is quite the norm, but, if I am right, one day it will be the province of cranks.

[288] While Mayr refers to "*seemingly* goal-directed behavior" (1974, emphasis added), he also shows that he takes goals to be real in some emergent sense when he says, for instance, "Even though there are indeed many organic processes and activities that are clearly goal-directed, there is no need to involve supernatural forces, because the goal is already coded in the program which directs these activities." (Mayr 1992).

G. Truly Illusory Goals

Attributing intention to behaviour is a primitive move. Children do it with no instruction, as when listening to nursery stories and playing. It is done in an instant, without conscious inference. If non-linguistic judgments can be countenanced, animals may also be said to make judgments of intentionality, as when prey distinguish pursuit from loitering in a predator.

—Lowell Nissen (1993)

Hence, it came about that, as Man awoke to the objective phenomena of directive correlation in nature, he began to describe them in terms of concepts introspectively derived from his own experience, and to interpret them in terms of anthropomorphic and psychological analogies—in terms of "purposes", "goals", "aims", &c. The resulting confusions in biology were fatal.

—Gerd Sommerhoff (1950)

As we have seen, no one denies that we commonly *perceive* goal-directedness; they only deny that biological goal-directedness is rooted truly in the natural world rather than being constructed somehow through the processes of perception. Before I argue one more time against eliminativism in the next section, I'd like to do something that eliminativists generally neglect to do in support of their positions: I'd like to highlight how illusion *can* convincingly account for the perception (and projection) of goal-directedness in certain circumstances where there is none.

Around a half century ago, Albert Michotte performed an extensive series of experiments on "the perception of causality" (1946/1963, 1950, 1968; see also Csibra 2007, 2008; Gao *et al.* 2009;

Heider and Simmel 1944), from which an interesting illusion emerged: Participants in his experiments observed feelings, thoughts, desires and goals in situations where there clearly were none. Here's how that happened.

Michotte was interested in the ways that people infer causation from motion, and so he arranged a set of experiments in which he could vary a number of parameters of motion and could then observe how those variations influenced viewers' perceptions of causality. One of Michotte's apparatuses displayed colored rectangles that appeared to move across a screen in various, simple, coordinated ways. The design of this apparatus consisted of a painted disc that was made to rotate smoothly and silently behind a screen with a slit in it. Curved stripes on the disc would then be seen, through the slit, as nearly rectangular shapes that would appear to move to the left or right, as the disc rotated. One can see that, were we to rotate the disc in Figure 7.6a, the two rectangles showing through the slit would at first be stationary, then the orange one would move right to abut the blue one for a moment, after which the blue one would move away before they would both remain stationary again. If this happened rather slowly, the description I just gave—in terms of rectangles moving independently—might be the way most people would conceive of what they saw.

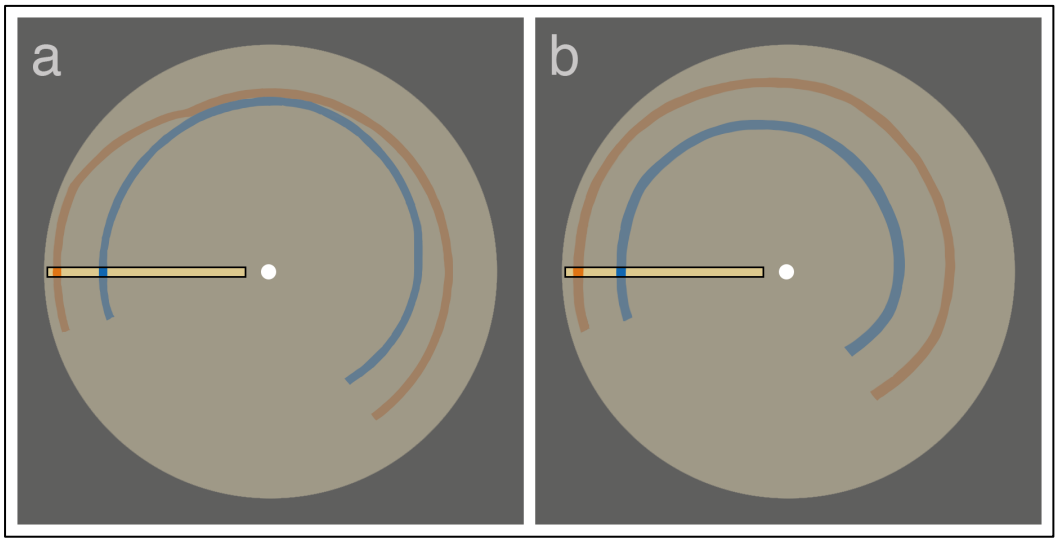


Figure 7.6: Schematics of display wheels like those used in Michotte's experimental apparatus. In both cases, the majority of the wheel would be fully occluded with only a small amount showing through the horizontal slit (here the majority of the images are only slightly greyed so that we can see what lies behind). When turned counter-clockwise, the colored stripes on wheel a would appear as rectangles that would move to the right, with the orange one moving first until it abutted the blue one, and then, after a pause, the blue one would move away. The stripes on wheel b would appear as the same two rectangles, both moving at the same speed and never touching one another. Of course, today, the same effect can be achieved much more simply on a computer screen.

If the period of abutment were reduced significantly (by repainting the disc otherwise) participants would observe a kind of billiard-ball causation in which the darker square *struck* the lighter one, knocking it forward. With subtle differences in the arrangement, the velocities, and the relative timing of the rectangles' movement, Michotte was able to elicit from his participants widely varying descriptions of the behavior of these rectangles. Sometimes people would describe the rectangles as moving left or right, or speeding up or down, or getting closer together or further apart; but other times they would describe them as "pushing", "launching", "fleeing", and so on. Much like the language in Beckner's examples that we looked at in the previous section, these

intentional verbs used by Michotte's participants all imply a perception of goal-directedness—an inference about a psychological or biological nature in the rectangles.

One particularly relevant and simple example here is the condition in which Michotte arranged for two rectangles to move together across the screen at the same velocity, one in front of the other. A disc like the one in Figure 7.6b rotating at a roughly constant rate would produce an image showing this kind of behavior. Michotte found that when this arrangement was presented at particular speeds, the rectangles would frequently be described as chasing and fleeing one another.

Of course they are doing no such thing. Rectangles don't chase or flee[289]. They move in lockstep simply because they were painted as two curves with a fixed gap between them. What, then, accounts for the use of goal-oriented language in describing their behavior? Or, as Michotte asks, "Why this tendency to translate the phenomena into terms of human or animal conduct?" (1968)[290].

The answer seems to be that there is a psychological predisposition in humans to perceive intentional, goal-oriented behavior. Psychology and philosophy of mind have given at least two names to this tendency. One is the "theory of mind", a term coined by Premack and Woodruff (1978) when investigating the extent to which chimpanzees have the same ability. The other is "the intentional stance", a name devised by Dennett (1987) in order to highlight how we approach the world from this "stance", even if unconsciously, in order to make successful predictions in a world that, for intelligent social creatures like us, is filled with intentional (thinking), goal-oriented friends and foes. Gergely and Csibra (2003; see also Csibra *et al.* 1999; Gergely *et al.* 1995) have amended Dennett's notion by suggesting that, prior to a fully developed intentional stance, human infants

[289] And, really, they are not even rectangles! They are just the illusions of rectangles made by viewing a portion of a curved line through a slit.

[290] A similar effect occurs with Braitenberg's notorious *Vehicles* (Braitenberg 1984). These simple robots only perform basic single- or double-rule behaviors such as approaching or avoiding light based on a single sensor or other such things. Yet most observers seem to think they are best described in terms of liking, wanting, fearing and other such intentional terms. Despite their simplicity, the "vehicles" seem lifelike.

show a "teleological stance" from which they are able to predict goal-directed, but not intentional (thought-based) behavior.

Another way of interpreting this is that the human mind has a tendency to project an image of intentionality or goal-directedness out onto objects in the world that resemble agents in some way, just as we have the tendency to complete any partial pattern we observe with the most likely candidate in our minds (see Figure 7.7)[291]. I have previously called this kind of tendency by the name "the projection error"; but of course it is an error only when performed zealously, since the projection of completed patterns is a normal tendency of perception, and a correct assumption when the partial pattern we perceived is a fragment of a real complete pattern (Hurley, Dennett and Adams 2011). In the case of the intentional stance or teleological stance, the patterns that are completed by the mind's automatic analogy mechanisms and then projected out onto objects in the world are the patterns of beliefs, desires, or goal-directedness, despite whether the object truly has any of these things[292].

[291] When viewing optical illusions such as those in Figure 7.7, our minds have a tendency to complete parts of the picture that are only fragmentary. This phenomenon of pattern completion is sometimes called "reification", a term that comes from the psychology of Gestalt perception and that means "making real". In terms of these illusions, certain patterns that are not completely present in the image, such as the triangle in 7.7A, are made "real" by the mind. A related concept is the Gestalt Law of Closure, which states that we have a tendency to perceive things as whole, closed forms—a notion that also explains, for instance, the reification of the triangle in 7.7A and the sphere in 7.7C as wholes, despite our only seeing fragmentary edges of them (Wertheimer 1923/1938; Koffka 1935). Elsewhere in perceptual psychology the term "filling-in" is used to refer to the way the perceptual machinery or the mind fills in the information that is not presented in the world, when building a representation of it (*e.g.*, Ramachandran and Gregory 1991).

[292] It is worth noting that while the terms "the intentional stance" and "the theory of mind" may sound as if they designate a special-purpose cognitive tool, if they are seen instead in the way I've described, then they are just one aspect of the normal pattern-recognition and analogical processes of cognition as applied to a particular type of content in the world (intentional content).

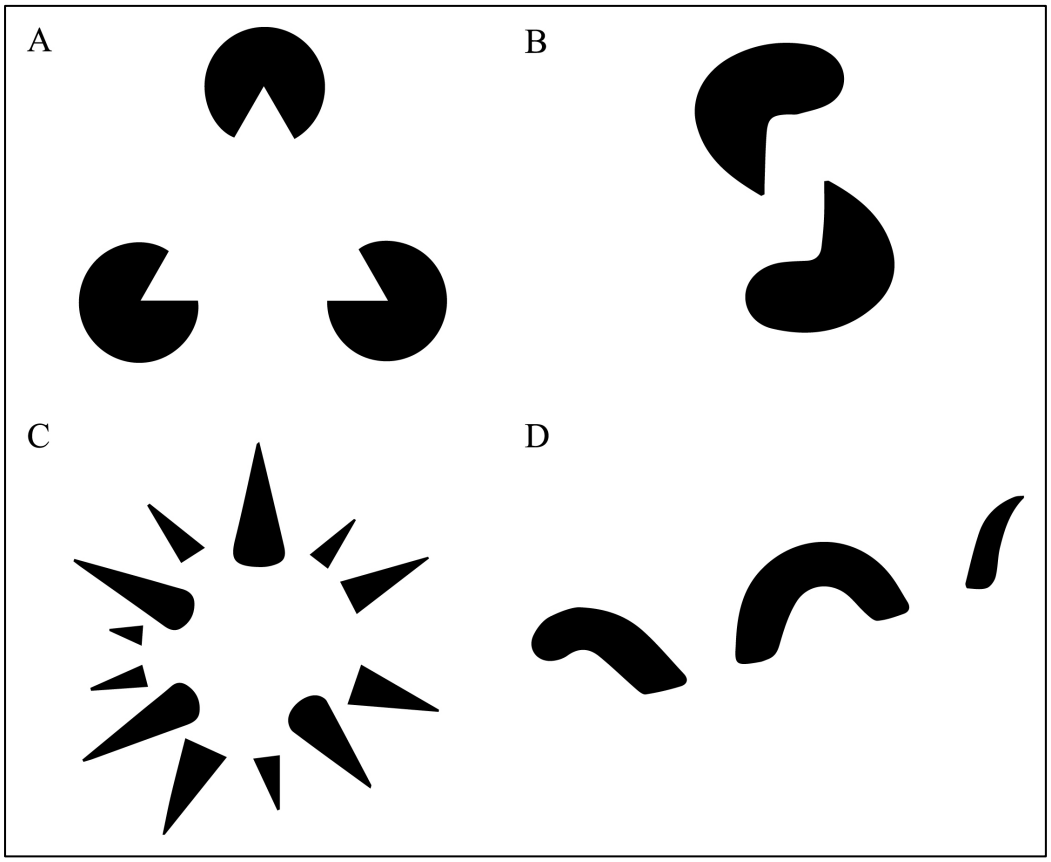


Figure 7.7: Gestalt perception illusions. When viewing images such as these we feel as if we "see" objects that are not truly present in the image. We see the triangle and sphere in A and C, and we complete the worms in B and D, coming to believe that there's a pole or a lake obscuring the rest of the serpent. Projecting such perceptions out into the world is a common task for our brains and usually it is correct. When we see forms that look like D in the real world, it is usually because there is a continuous body of, say, a Loch Ness monster half-submerged in the lake, and so imagining the rest of the creature is the right thing to do. But in some less common cases, that projection will turn out to be false. And in drawings such as these ones, meant to be ambiguous, there may not be a truth to the matter since there is neither a lake nor a Loch Ness monster on the page—there are only some black marks against a white field.

So the teleological stance or intentional stance, working in the minds of observers, is how we come to see Michotte's rectangles as wanting to do certain things, and how we come to see them not only

to be *moving*, but to be "chasing" or "fleeing". As long as the motion of the rectangles is sufficiently similar to actual cases of chasing or fleeing that we've experienced in our lives and sufficiently different from most of the movements of bodies normally influenced only by mechanics—as long as the motion occurs at certain rates, directions, and relative timing and distance—we can draw the analogy effortlessly and see the rectangles as being lifelike and as having goals.

H. Limits to the Illusion

The main question . . . is not whether the processes of life can properly be called purposive: it is rather the question if the purposiveness in those processes is the result of a special constellation of factors known already to the sciences of the inorganic, or if it is the result of an autonomy peculiar to the processes themselves. For that there is, as a matter of fact, much that is purposive in vital phenomena is merely an immediate deduction from the definition of the concept of purpose itself, and from the application of this definition to living beings.

—Hans Driesch (1914)

As we saw earlier, the goal eliminativist wants to extend the scope of the illusion just discussed. The eliminativist would suggest that perhaps the sunflower or squirrel or bacterium is very much the same as one of Michotte's rectangles. Perhaps when a sunflower moves towards the sun it is only *moving towards* the sun, not following or tracking. Perhaps the sense we get that the bloom might be in pursuit of the sun is due to the illusory outward projection of goal-directedness, caused by the tendency of our minds to take the teleological stance.

To add a little more fuel to the eliminativist fire, we might note that the psychological projection of goal-directedness is hardly limited to rectangles on screens in laboratory situations. When we reviewed animism, we saw a number of other versions of this illusion. People often conjure spirits, ghosts, and gods or other willful, goal-directed activity to explain the behaviors of such things as volcanoes, storms, tsunamis, earthquakes, obstinate nuts and bolts, statically charged hairs, doors shut by breezes, and so on. The illusion is widespread. So why shouldn't we expect similarly overzealous attributions of goals to be responsible also for our perceptions that bacteria, plants, and animals are purposive?

The difference is simple. Any time we are dealing with an illusion, if we search long enough and carefully enough, we can determine that what we see is an illusion either by varying the conditions of our perceptual interaction with the phenomenon until it disappears (in which case we know the phenomenon is an illusion but don't necessarily know what causes it to manifest) or by finding an alternative explanation that perhaps forestalls the illusion from occurring or at least makes sense of the illusion in terms of our perceptual machinery or viewing perspective.

In the case of Michotte's rectangles, a participant in the study needs only to get out of their chair and investigate the apparatus to discover that they are not really viewing rectangular objects at all (much less goal-directed rectangles). As soon as they probe the device, their illusory impressions of those rectangles "chasing" and "fleeing" will evaporate entirely.

If we believe that stubborn statically-charged hairs are willful, we only have to observe them in a broader array of circumstances to eventually discover that they usually do nothing of the willful sort, and that they only stand up seemingly obstinately when they've been treated in a certain way—say, when a wool sweater has just been dragged across them on a dry day. As soon as we find our perception of the hair's obstinacy to be regularly associated only with certain circumstances, then we will suspect that it is an illusion and that something other than goal-directedness might account for the hair's behavior. By further investigating those circumstances, we eventually may find an alternative explanation for the hair's behavior (say, in terms of electrical charge) causing the illusion of goal-directedness to be unconvincing, even if we still seem to perceive it.

Similarly, the reason we no longer believe goal-directedness to play a role in geological and meteorological events is that, over time, science has found alternative explanations for most such phenomena. Because those alternatives are convincing, they have made the illusion of goal-directedness disappear. This is how, bit by bit, materialism came to dominate in physics, while ideas of cosmic and theological teleology had their credibility gradually eroded.

Now the fact that no one yet has found an alternative explanation for organismic goal-directedness is certainly not proof that such an alternative doesn't exist. We may not have searched long enough or broadly enough or carefully enough. There may still be some way to make the observed pattern of goal-directedness in organisms flicker in and out of existence by changing our perceptual point of view . . . there may be some context into which we can drag organisms thereby rendering their behavior suddenly objective rather than projective, such that nothing appears to be good or bad for them, such that their activities no longer look like strivings, and such that our intuitive sense when viewing their behaviors is that those behaviors are not *for* anything. But the challenge, in order for the goal-eliminativist to support their thesis, is to *find that context* and to show us that they can reliably erase our perception of goal-directedness from the organismic world (and reintroduce it again) at will, by changing the relationship of organisms to the perceptual system (and of course without changing the organisms themselves). Until that can be done, the logical assumption to make is that the pattern of goal-directedness that we all observe in every living organism, from bacteria to humans, is real.

In the next two chapters I will describe a hypothesis about the nature of goal-directedness, synthesized from the ideas of many modern theoretical biologists. Rather than trying to avoid the horns of the teleologist's dilemma, this synthesis follows the lead of post-behaviorist psychologists and philosophers and instead embraces a version of the second horn—accounting for subjectivity in objective material terms, and thus showing how an object can truly become a subject.

Part II

Chapter VIII

Natural Purposes

What is the characteristic feature of life? When is a piece of matter said to be alive? When it goes on 'doing something', moving, exchanging material with its environment, and so forth, and that for a much longer period than we would expect of an inanimate piece of matter to 'keep going' under similar circumstances.

—Erwin Schrödinger (1944)

At this point, I am going to try to develop an account of what it means for a pattern to display vitalistic, projective, teleological tendencies—for an animate piece of matter to go on "doing something . . . for a much longer period than we would expect of an inanimate piece of matter". The account I will present is a new twist on an older account, the larger fragments of which began to appear during the 1970s. The account is still far from complete, even with the amendments I offer here, but the ways in which it neatly ties together so many of the otherwise philosophically elusive concepts that we have been discussing has given me the confidence to believe it is on the right track.

Aside from Immanuel Kant, who provided the most central contribution over two centuries ago, the major contributors thus far are Manfred Eigen, Tibor Gánti, Stuart Kauffman, and Humberto Maturana and Francisco Varela (all of whom have further developed *the Kantian notion of reflexive self-organization*); Kurt Gödel, W.V.O Quine, and Douglas Hofstadter (the latter of who has picturesquely discussed and made clear the work of the former two, highlighting *the dual roles a self-replicating structure can play both as functional mechanism and as informational payload*); and Richard Dawkins

(who has recast our conception of organisms from being individuals to being *collectives of economically cooperative and competitive beneficiaries*).

The modern theory of teleology is an attempt to weave the themes above into a single cohesive story. That story claims that basic teleological agents are patterns of matter whose structures, in serving as both mechanism and payload (Hofstadter), are able to reflexively maintain their own organization (Kant, Eigen, Gánti, Kauffman, Maturana and Varela). The story gets to be even more interesting when we see that these structures may become organizationally coupled by way of economic relations to form larger, materially overlapping collectives that share parts and processes in service of both their individual and common good (Dawkins).

My contribution to telling that story will be twofold. First, I will work to show how these thinkers' ideas may be fitted together into this unified account. And second, I will attempt to use, as an adhesive in fitting those pieces together, the beginnings of a new quantifiable model of the fundaments of subjectivity. That model will be incomplete, but, as I said, I think it will have a variety of interesting consequences, and I think it will represent a significant step away from a number of philosophical impasses, including Theseus' ship, many historical debates over normativity and, especially, teleological eliminativism and proper functions, and, at the same time, a step towards a new era of teleological realism and a science that may one day encompass subjective phenomena within its objective topics of study.

I will also take the time to explore the ways in which the pieces above seem to naturally produce a tripartite classification of the kinds of material orderliness that can exist. This comes with the territory that we'll be exploring because one of those three categories of orderliness will constitute the modern theory of naturally teleological patterns.

A. A Preview of the Theory

Pregnant in the birth of the universe was the birth of life. Agency may be coextensive with life. Life certainly burgeons nowhere without agency. We all act on our own behalf. In the Kantian form: What must something be such that it can act on its own behalf?

—Stuart Kauffman (2000)

I am first going to make a rough sketch of the way I see the pieces of the modern theory of teleology fitting together, and then, in the next few chapters, I will enrich the picture with bolder lines and colors that will help bring it to life. We can assess the theoretical value of the offering, both descriptively and predictively, as a later exercise.

The theory of teleology is not about the kinds of goals we typically think of. It is not about reaching the sales targets for the quarter or growing enough rice to feed the village, and it is not about squirreling away nuts for the winter or swimming upstream in search of greener pastures, although at the end of the day, in one manner or another, it should figure prominently in an explanation of all these things. The situations just described can all be considered to be *subordinate* goals—some kind of partial contributions to an ultimate goal in a hierarchy of ends. But there can be no subordinate goals without the *ultimate* goals they contribute to.

At its root, the modern theory of teleology is about those ultimate goals, which, the theory claims, come in only one kind: the goal of continued existence, for some organizational pattern, in the face of a relentless swarm of attacks led by the material decay of ratcheted braising. To apply a more concise term, the ultimate goal in this world is nothing other than the *persistence* of a pattern.

Something that is goal-directed is something that has a set of behaviors, which, taken together, give rise to its own existential persistence.

We can differentiate the teleological brand of existential persistence from other kinds of stable persistence, such as the long-term viability of stable atomic isotopes or the spontaneously reoccurring stability of structures such as lipid micelles or Bénard cells. In particular, the stability found in the teleological version is driven in part by the activities or motions of the persisting pattern itself, and because of this, we can term it "active stability" or "active persistence". A goal-directed pattern persists through a combination of external factors and, importantly, its own *actions* (see also Juarrero 1999, who analyzes action and agency in similar terms; and Pross 2005, 2008, 2009, who labels a similarly active notion of persistence at the level of chemical replicators "kinetic stability").

But already we seem to have made a bit of a magical leap here when we transition from talking about an item's *motions,* to talking about its *actions*; the former is a relative but nonetheless objective notion while the latter is an agentive and thus fundamentally subjective one. The best way I've found to describe why active persistence underlies this magical transition from the objective to the subjective is that, when a mere, objective, physical pattern of forces undergoes a group of mere, objective, physical motions that just happen to have the reflexive property of facilitating the persistence of that very same pattern, it thereby comes to be doing what can best be described as *helping itself* (see Kauffman 2000). And in the two parts of that short phrase—"help" and "itself"—we can find the roots both of benefit and of identity, the fundaments of subjectivity, the germs from which the rest of the projective and agentive features in the universe emerge, grow, and thrive.

The modern theory of teleology is of course a theory of goal-directedness but, really, the terms it is given in comprise a theory of what it means to have a self and of what it means to be able to benefit, to evaluate, and to act on one's own behalf. When a thing comes to be able to actively

persist, the great leap is made from a prior world in which "Why is it doing that?" could be answered only with a *process narrative*, to a new world in which that question can in some cases be answered with a *reason*—a world in which the "it" that "is doing that" becomes sharply defined rather than observer-dependent, and in which the "why" begins to refer to "what for" rather than "how come". The closing of the autocausal circle (in which a thing is able to help itself) creates a distinctive situation in which each behavior in the circle is not only the cause of the other behaviors but also, transitively through those other behaviors, the cause of itself, and so the previously unanswerable, future-directed question "*What* is it doing that *for*?" suddenly comes to have a meaning that is rich enough to replace the purely historical "*How* has it *come* to be doing that?" The answer, which can for the first time be given in teleological terms (using phrases such as "for" or "in order to"), is simple: It is "doing that" (whatever it is that it does) *in order that it may persist*.

Identity

The way the theory explains identity is perhaps the most complicated piece but it is not too difficult to understand, taken a step at a time. Most fundamentally, it claims that an individual teleological identity is only the persisting pattern itself—the set of capacities that are, by their actions and their mutual interactions, able to create themselves (Maturana and Varela 1973).

This is an identity for the simplest of reasons: since this set of capacities comprises precisely the pieces required to create those same capacities, it therefore remains *identical* across time and thus persists for longer than its disconnected parts otherwise would. But this short formulation requires quite a bit of interpretation. For one thing, what it means to be "identical across time" needs to be made precise. We will need to know how to interpret different versions of an ever-changing pattern as being *the very same thing*. For another thing, we also need some way to provide precision to the

concept of "longer than . . . otherwise"—just how long would "otherwise" have been? For a third thing, we need to specify what the term "create" means, and what the background context is with respect to which that type of creation must be understood. What are the physical requirements and constraints for this notion of "creation"? Just what is being created? Obviously it cannot be *the atoms* from which the pattern is made (they are, more or less, enduring); instead, it must be something about *their organization*. And for a fourth thing, we should recognize that the ways in which we think about these previous three concerns will ultimately shape a new concept of identity that will no longer be rooted in intuitively concrete notions such as objects or items or bodies. We need to be prepared to accept the fact that this new type of identity will integrate differently with many other concepts that compose our current scientific and philosophical understandings of the world. I'll try to address each of these pieces throughout this second part of the dissertation; for now, let's continue our preview with an introduction to how the persistence of identities can underlie a useful, if relative, notion of value.

Value

A cyanobacteria colony in a sun-warmed patch of the ocean may persist for months, and its clonal progeny may ultimately persist for billions of years. In contrast, an influenza virion or, for that matter, silver halides or even diatomic oxygen, in the same direct summer sunshine may last only moments before decomposing into smaller pieces. Success or failure at persistence is the archetypal standard that patterns that may come to exist must live up to; merely by *being there* (in a particular environment), a pattern either is the kind of thing that persists (thus it passes the test and succeeds) or it is the kind of thing that does not (thus it fails). *Every pattern in the world is implicitly held to this natural, objectively normative standard.* And because of this, the possibility of persistence, the

potential to persist, by being something that is allowable by the physics (or the mathematics) of our world, ushers in a naturally emergent transition from a universe of *objects* to a universe of *subjects* . . . a universe in which evaluatively normative patterns—those that succeed at persistence—may potentially come to exist.

This brand of normativity is *evaluative* rather than *comparative* because the "ought" (when we say that a candidate pattern "ought to persist") is measured in terms of success or failure, and not in terms of any comparison against a distribution of other patterns in the world (see Chapter I). We are saying "this thing ought to persist" not because it measures up to a comparison class, nor because it is like other things that persist (although that might also be the case), but only because its behaviors, if it works properly, encourage its own persistence. If a thing has properties that engender its own persistence in a certain environment, then it ought to persist in that environment; and if it does not have such properties, then it ought not to persist.

And since persistence gives birth to an evaluative normativity, it also becomes the underlying standard against which any other items and events in the world may come to be evaluable. Objects and events can be evaluated in terms of their relative contributions to the persistence of identities. A thing is "good" (for any particular persisting pattern) if it helps that pattern achieve its goal (of persistence) and it is "bad" (for that pattern) if it thwarts the pattern's efforts towards achieving that goal. Of course, if a thing neither helps nor hinders the pattern's efforts, then it is evaluatively neutral. Evaluation, then, straightforwardly gives rise to value and thus, in the presence of other evaluating identities, economy: If a thing is good for me, then I consider it to be good. And if *I* consider it to be good and *you* consider it to be good, then, depending on various factors (such as whether it is sharable or limited), we may cooperate or compete with respect to it . . . And so, hand in hand with the emergence of persistence in our world comes the emergence of evaluative (and

subjective) notions such as being *useful*, working *properly*, having *value*, being *good for*, being *mutually beneficial*, and also, ultimately, being *ethical*.[293]

I have made mention of the concept of economy, aware of the fact that it sounds as if my cart is now getting ahead of my horse. In this piece of work, I don't intend to analyze the implications of the theory of teleology for higher-level economic theory, though I think the bases of that extrapolation will become clear in short order[294]. Rather, I mean only to set the foundations for understanding the economic principles that are able to link smaller teleological identities into larger teleological identities, by way of a blend of competitive and cooperative relationships. In short, these are the very same already-well-studied principles that underlie not just biological parasitism, mutualism, and commensalism, but also the social and commercial versions of the same phenomena.

Adding it All Up

To recapitulate this brief introduction, then: The short form of the modern theory of teleology holds that patterns that have methods by which they might actively persist, despite experiencing the inexorable material decay caused by ratcheted braising, *thereby* gain identities and thus bring into existence the relative, subjective evaluation of objects and events with respect to the goals of maintaining those identities. A pattern that is able to help itself persist is a pattern whose actions are jointly goal-directed towards that persistence[295], and whose parts serve the purpose of aiding in those actions. This is how the subjective phenomena of *identity*, *value*, *purpose,* and *goal-*

[293] For the philosopher: I suggest that the description in the previous two paragraphs shows precisely how we can "get an ought from an is", thus patching back together the positive and normative worlds that "Hume's Guillotine" once carved from one another (Hume 1738). This then provides a basis for naturalistic treatment of moral philosophy and secular ethics. What one then ought to do is whatever is good for one's own identity (since one's fundamental duty is persistence); however, what is good for one's identity is a proposition that can only be clearly understood in terms of what constitutes one's identity, and, especially in the case of humans, that turns out to be a very, very complex thing.
[294] For one thing, much of modern economic theory depends on the assumption that agents—often, rational agents—exist. The current work provides a theory of agency that can underpin those assumptions.
[295] And thus it is also a pattern whose behaviors can rightly be called *actions* (see also Juarrero 1999).

directedness all get their footing in our objective world, and provide a basis upon which further subjective phenomena may be built. Compositions of these teleological identities, in both time and space, form the basis for the kinds of agencies we normally call *organisms* and *species* and, eventually, *societies* and *governments* and so on. When we've got a natural theory of goal-directedness under our belts, the biophilosopher's notion of *functioning* can easily be understood in relation to it, in more or less the way that Christopher Boorse (1976, 2002) has advocated: Anything that (verb-) functions does so because it plays a causal role in the persistence of some teleological identity. In order to be seen as having a noun-function (even if that vision is illusory), a thing simply needs to verb-function in a particular way with some regularity, or to have been constructed solely in order to potentially (but not necessarily) verb-function in a particular way at some point in time.

Coming up next, we'll spend some time analyzing the notion of persistence itself in order to differentiate the most general methods by which a thing might persist. The key notion to keep our eye out for will be *informational redundancy*; we are looking, here, for nature's blueprints—the sources of information that make the creation and maintenance of orderliness possible. After that, in the section on "autocausality", we'll look at the provocative yet presently vague method of being-both-cause-and-effect-of-oneself as a means of realizing persistence. There have been a number of specific models of autocausality offered by theorists, and we'll briefly review the most prominent of them, because the more generic offering we are developing will maintain some relationship with each of them. In the next two chapters, we'll sharpen the notion of autocausal persistence by presenting a theory of identity based on informational redundancy and given also in terms compatible with the materialist view of causation. Once we've got a handle on how certain things may remain retain an identity in the world, we'll then differentiate between those that retain their identity by helping themselves, and those that retain their identity under the protection of other processes. Both may have identities, and both may potentially benefit, but only the former benefit

in part through their own activities, and thus only they can be said to have agentive actions that are goal-directed towards that very benefit.

B. Persistence

To be or not to be, that is the question.

—Hamlet (3.1.64)

The question of life or death seems to be the most crucial concern for any biological entity, but roughly the same idea arises even pre-biologically: Along with any physicochemical structure that, by whatever method, *comes to* exist in the world, also comes the potential question of whether or not that structure will *continue to* exist. The theory we are going to look at is not so much a theory of the origins of various organizational patterns as it is a theory of the existential destiny of those patterns.

The central suggestion I will make is that teleological patterns, because of an information-theoretic feature of their architecture, form a class with a qualitatively different existential destiny from that of other major categories of patterns. Teleological patterns are neither guaranteed to persist, nor guaranteed to fall to pieces. Quite literally, there is a sense in which their destiny is in their own (figurative) hands.

The Struggle for Existence / A Recipe for Persistence

The importance of persistence as a hallmark of the living has been noticed by any one of us who has woken up and killed a beast before breakfast; but the topic really gained theoretical prominence when Darwin (1859) borrowed a term he found in Thomas Malthus' (1826) *Essay on the Principle of Population.* The key idea was what both writers referred to as the "struggle for

existence"—the fact that an individual organism must struggle constantly against the causes of mortality in order to continue existing (see also, *e.g.*, Wallace 1858; Huxley 1863; Weismann 1909).

Darwin made use of Malthus' idea primarily to describe the competitions that take place between an organism and two classes of resource-hungry competitors: its peers and its predators. In Darwin's sense, it is the winners of both these competitions that tend to persist, surviving not only until tomorrow but also, one hopes, long enough to be represented in the next generation.

Following on the heels of Darwin's theory, the biologist Herbert Spencer coined the now-famous term "survival of the fittest", underscoring the Darwinian characterization of organisms as kinds of persistors (Spencer 1864; see also Darwin 1868, 1869).[296] Nearly a century later the psychologist and AI pioneer Herbert Simon riffed on Spencer's expression, noting that it isolates a special case of a more general existential bias in the universe: the relatively greater persistence of the relatively more effective persistors, or, perhaps more succinctly, *the persistence of the "stablest"* (Simon 1962; see also Bouchard 2004; Dawkins 1976; Godfrey-Smith 2009; Keller 2007; Rosenberg & Kaplan 2005; Rosenberg 2006).[297]

From Simon's perspective, the struggle for existence should really be seen not as a struggle against other individuals or other species, and not even as a struggle fought only by organisms, but merely as the struggle of *any potential pattern* in the world against changes to its material organization. Predation and peer competition are just some of the more salient cases of what organisms may

[296] Framing natural selection in terms of *persistors* sounds a great deal like framing it in terms of *replicators* (Dawkins 1976, 1982; Dennett 1995; Haig 1997; and Hull 1980, 1988), and the reader attuned to that terminology will likely be reminded now of the complicated debate over the validity and usefulness of the replicator framework, including, especially, the definition of what exactly comprise the fundamental units of replication, and thus of selection. Whatever the case may be, even the classical framework for thinking about natural selection takes, as an unquestioned assumption, that there is a population of reproducing agents, and that this reproduction means that *something*—whether or not that something is well defined—is persisting in some way over time.

[297] This is an idea that is easier to see as being clearly true in its abstract form than it is when applied to real-world biological structures. Although we will try to address this issue with more clarity later, in biology it is not always clear just what persists across a reproductive cycle. Since most organisms are sexual reproducers rather than clonal organisms, their progeny share genes and traits that come from each of two parents, and neither parent has persisted alone. For now, I'll continue to pursue the abstract notion.

struggle against; one also struggles for existence against famine, against drought, against solar radiation, against falling rocks, against storms and high seas, against cancer,[298] against stroke, against neurodegeneracy, against dehydration, against environmental toxins, and quite generally against the depredations of time, none of which are necessarily caused by competition with other organisms.

This formulation lacks detail (about the causes and modes of stability and persistence), but we shouldn't underestimate its importance. What Simon's idea suggests is that *the grandest ontological categories of our world—the kinds of things that we can expect to find existing in any universe where braising takes place—are those that have reliable methods of persistence.* This is an idea that paves the way for framing our ontological inquiries about what philosophers call the furnishings of our world. So the main waypoint, then, on the way to "What kinds of patterns could possibly exist?" is the following question: "What modes of persistence might there be?"

Resilience vs. Redundancy

In looking for modes of stability against the inevitably accumulative damage of braising, one might imagine three abstract possibilities. First, we know from biological examples that a thing might have some method by which it could alter its own fate by protecting or repairing itself. Second, we understand that a thing might fall under the protection of some benefactor processes that are disposed to intervene on its behalf, shielding it or patching it up as necessary, the way, for

[298] Perhaps it is debatable whether cancer can be thought of as another organism. After all, cancers are by definition genetically distinct from the normal cells of the host, and so one might consider them to be different from the host. While that may seem like a blurry distinction to make, certain cases can be even more pronounced. For instance, there is the case of facial tumor disease in Tasmanian devils—a contagious parasitic cancer that developed originally in one individual, but has become highly transmissible between hosts, and now threatens the species with extinction. In this case, the disease is an infectious rather than developmental disease; it has its own genome that differs from that of its hosts, and it conducts its own reproductive lifecycle quite analogously to any other distinct obligate parasite (Pearse and Swift 2006).

instance, we humans repair and maintain many of our artifacts. And third, a thing might simply be tough as nails.

The first two manners are worth our consideration, but let's take a few moments to dismiss the third possibility. Any sort of resilience—the hardness of diamond, the durability of stainless steel, the tensile strength of spider's webbing—can only be measured relative to the distribution of braising energy that a pattern might encounter. If I build a stronger shield, you can always build a stronger weapon. And, as physicists have determined, everything is eventually susceptible to dissociation under extreme enough conditions.

A pair of examples can help in directing our attention further away from resilience and instead toward the role that redundancy plays in the first two manners of persistence. The examples we'll use are an expanded-polystyrene foam model of a shark and a real adult Greenland shark. The grounds on which the two of these sharks are comparable is that, unlike most other relatively soft structures we know of, both of these usually can retain the bulk of their relatively soft shape and structure for hundreds of years. The comparison becomes interesting because the equally substantial longevity in each of the two cases must be accounted for and understood quite differently.

In the case of the polystyrene foam, while the shark's structure changes only very slowly[299], it is nonetheless on a one-way trajectory along which it transforms, bit by bit, into various alternative structures. The slings and arrows that it suffers will result in a slow, but irreversible accumulation of bruises and scratches, much as we found to be the case with the washing machine that we flung through an asteroid field in Chapter II.

The Greenland shark's flesh is also heir to a thousand natural shocks, but there are two differences in its constitution. First, it is much more fragile and thus susceptible to more rapid

[299] For a long time, polystyrene foam was thought to be entirely resistant to biodegradation. However, recent research has found an exception to this rule in a bacterium that lives in the guts of mealworms (see *e.g.* Yang *et al.* 2015).

changes.[300] But second, changes to the living shark are not uniformly destructive of its organization; for most of its life there is as much constructive change as there is destructive change.

We can look more closely at these differences in terms of the informational content of each shark. The polystyrene shark has no informational redundancy—its structure is specified precisely once, by its very existence. As its body gets banged about by the vagaries of existence, material disorder in it increases—pieces get increasingly bent out of shape, and some may fall off entirely. Because there was only ever one copy of the informational specification of that body, any damage represents the irretrievable loss of information about its material organization. One could only reconstruct a bruised polystyrene model by making use of *external* information—by using a manufacturer's blueprint of some sort or by making educated guesses the same way a restoration artist might repair renaissance paintings or ancient frescoes. One could assume that a damaged patch had once been smoother or that a lost fin had once had a particular shape, but without any detailed documentation (beyond the shark's own now-damaged form) one could never be sure, and there is always a risk that one might unknowingly botch the job.

In the Greenland shark, any structural information that is lost to braising is also gone. However, there is redundant information about the shark's material organization preserved elsewhere in the functional structures and processes of the shark's undamaged parts (especially but not solely in the DNA). There is a blueprint of a sort, inside the shark, and it is that redundant informational content that lies at the heart of the reconstructive processes that prevent the accumulation of decomposition in the shark. Those vital processes "know" how to repair and rebuild the degraded parts of the shark because they are informed by the redundant information contained in the blueprint.

[300] At present, the Greenland Shark is known to be the longest-living vertebrate; one study's authors have estimated a recent specimen to have had a lifespan of nearly four hundred (plus or minus a broad error margin of one hundred and twenty) years (Nielsen *et al.* 2016).

As we continue our explorations, we are going to find that, at the heart of the process of active persistence by which a pattern's own activities contribute to its persistence, there always lies the notion of *redundant information*. This partly explains why the Greenland shark—laced with redundant information throughout its body—is able to recover from damage and disease, and why the polystyrene shark—comparatively devoid of redundant information—may be highly resilient but can never heal a wound.

C. Autocausality

It must be thought of as an organ that produces the other parts (consequently each produces the others reciprocally), which cannot be the case in any instrument of art, but only of nature, which provides all the matter for instruments (even those of art): only then and on that account can such a product, as an organized and self-organizing being, be called a natural end.

—Immanuel Kant (1790)

If braising is unavoidable and resilience is not a realistic option, then would-be persistors seem to be left only with the first two strategies mentioned—a pattern can either help itself rebuild lost structure (like the Greenland shark and other organisms), or it can be the lucky beneficiary of external reconstructive help (like crystals and micelles or like a polystyrene shark under the care of a preservation league armed with a detailed blueprint). We'll explore both strategies, as each can produce a persistent identity. However our focus eventually will be on differentiating between these manners of persistence in order to isolate the particular organizational features that contribute to the teleological capacity by which an identity is not only able to be helped, but able to help itself.

To Be Cause and Effect of Oneself

Kant's contribution was the notion, quoted first in Chapter III and again in the epigraph above, that organisms are teleological because they are "self-organizing"—because they are both cause and effect of themselves. However, Kant's use of the term "self-organizing" differs somewhat from contemporary uses, and it will be worthwhile for us to notice the difference.

In modern scientific contexts, beginning with Ashby (1947), "self-organizing" has come to mean something more akin to spontaneously organizing. If we take, for instance, the cases of crystallization or the formation of micelles or Bénard cells, we have processes that, given the proper contextual conditions, are able to spontaneously produce materially organized structures (the crystals, the micelles, and the Bénard cells). However, it is important to note that, in each of these cases, the structures themselves need not play a role in their own creation. There is no need for a "self"—a prior version of the pattern—to exist before the organization comes to be created. Before the proper temperature gradient is reached in a fluid system (and without the proper viscosity and fluid layer thickness), Bénard cells simply do not exist. They organize spontaneously, they become organized, but they do not organize *themselves*. They—their own dynamical effects upon the world—play no necessary role in their coming into being. The same goes for crystals and micelles that spontaneously emerge from solutions when conditions are right, but which do not necessarily require prior selves as either blueprints or machinery used in the formation of later ones.[301]

In contrast, we can tell from Kant's discussion and from his phrase, "both cause and effect of itself," that he was using the term "self" in a stricter sense. Kant intended his "self-organizing" to refer to a process that is also *reflexive* rather than merely *spontaneous*, a process in which the contextual conditions required for the formation of a pattern notably include the pattern itself. Unlike crystals and Bénard cells, new bacteria, for instance, form only in conditions that include "old" bacteria; there are roles played by the internal machinery (and blueprints) of a bacterium that are necessary to the formation of new bacteria. As Pasteur showed decisively, while the notion of spontaneous generation can be used to describe the formation of crystals and micelles, the generation of

[301] These structures may also, at times, be perceived as "self-healing", a term that seems to imply an act in which the self plays some role. However, the term is a misnomer—the healing in these processes consists of the very same process as the initial formation, and it has no necessary dependency on the existence of a self—a prior complete version of the object that is being "healed". The same external conditions are all that is critical to both the ability for these structures to form and also for their damage to be repaired.

(organizationally-) complex living forms simply doesn't occur spontaneously. As Raspail and Virchow put it: *Omnis cellula e cellula*. All cells come from cells. Kant's self-organization is a qualitatively different (and as we will soon see, also quantitatively different) mode of behavior than Ashby's.

Early Models of Persistence

For whatever historical reasons, during the early 1970s the time apparently was ripe for theoretical biologists to expand upon Kant's notion and so, nearly simultaneously, a number of descriptions were offered of what it might mean for a thing to persist by being both cause and effect of itself. As we'll see, Tibor Gánti, Manfred Eigen, Stuart Kauffman, and Umberto Maturana and Francisco Varela each attempted in their own way to define a biological or pre-biological identity in terms of autocausality. In all cases, the autocausal systems suggested by these theorists are ones in which the structures involved have redundant causal potential that increases their own chances at persistence. Unfortunately, as they stand, the schemas that are precise are too specific, and autopoiesis—the one that is more general—is too vague.

Gánti's version was what he called a *chemoton*—a chemical automaton. A chemoton is essentially a stripped-down model of a cell that enumerates a minimal set of interconnected cell-biological capacities that appear to be cause and effect of one another. The original chemoton consists of 1) an autocatalytic metabolism that mobilizes free energy and material from the environment, 2) an information-replicating mechanism that uses the byproducts of the metabolic system to reproduce itself as well as to produce components used in the membrane production subsystem, and 3) a membrane production subsystem that manufactures molecules that self-assemble into an enclosing membrane. This membrane not only allows the ingress of the energy

sources that fuel the metabolic subsystem and the expulsion of wastes, but also keeps the machinery in the metabolic and synthetic subsystems all near enough to one another to operate and interoperate properly. The main idea is that these three high-level capacities all support one another in such a way that none of them can exist without the others. The three capacities are cause and effect of one another, and so, taken together as a whole, the chemoton is cause and effect of itself (Gánti 1971/2003, 1997; see also Deacon's, 2013, *autogen;* and Agmon's, 2015, *protocell*, for related models).

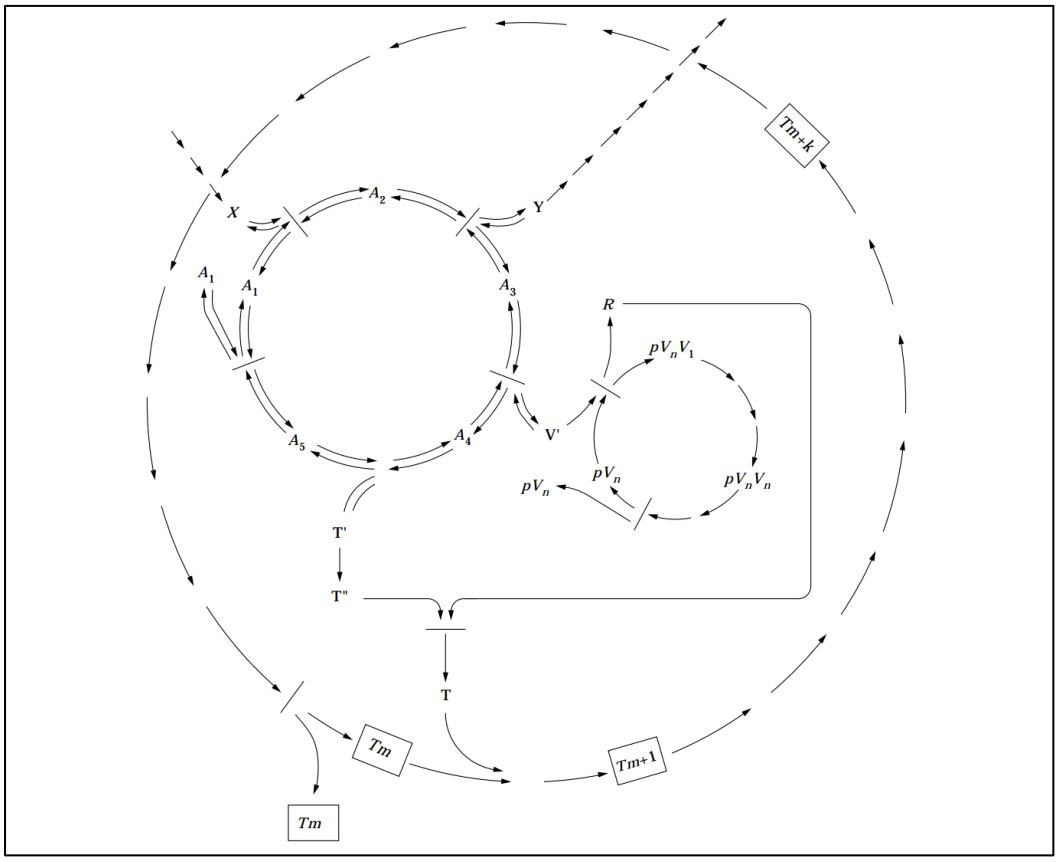


Figure 8.1: A diagram of a *chemoton* that elaborates upon Gánti's original (1971) model. The technical complexities of this diagram will be explored a bit later; what is important to the current discussion is only the general outline of its form. The metabolic subsystem (the circle in the upper left, made of arrows connecting various "A" labels) consumes nutrients X and produces wastes Y and, in the

process produces its own parts, as well as necessary precursors (V' and T') for the other two subsystems. The other small cycle, just right of center and composed of arrows connecting various "pV" labels, represents the "information" subsystem, which not only produces itself through a process of template replication but also results in a precursor component (R) that contributes to the membrane system. In order to do this, the information subsystem borrows energy and materials (symbolized by V') secured by the metabolic subsystem. The large enclosing circle composed of "Tm" labels represents the membrane production subsystem, which, using resources R and T", results in the membrane that allows the ingress of X and the expulsion of Y and helps maintain the functional proximity between the other parts. Most importantly, it is the interconnections between these three cycles that constitute the causal, catalytic coupling that Gánti suggests joins the three subsystems into a singular, living unity. Not only is this system meant to be self-sustaining, but, since each of the subsystems potentially duplicates its own members (*i.e.*, A_1 becomes $2A_1$; T_m becomes $2T_m$; and pV_n becomes $2pV_n$) it may also grow to contain redundant parts, providing the chance to replicate by binary fission. Reprinted from Gánti (1997).

Another version of "cause and effect of itself" is what Kauffman calls *collectively autocatalytic sets*—groups of chemical species, each member of which plays a role as a catalyst in the spontaneous generation of some other member or members of the set such that, all together, through these catalyzed synthesis reactions, the set produces a future version of itself that contains all the same members (Kauffman 1971b, 2000). So, for instance, if chemical species A catalyzes the synthesis of chemical species B, and if B similarly catalyzes the synthesis of C, and C does the same for A . . . and if there are enough raw materials floating around for all these processes to proceed . . . then A, B, and C, which are unable to spontaneously form in the absence of one another, may all come to be abundant in the presence of one another. Kauffman's work suggests that a biological organism, with its myriad interlinked enzymatic chemical processes, may be a highly complex autocatalytic set.

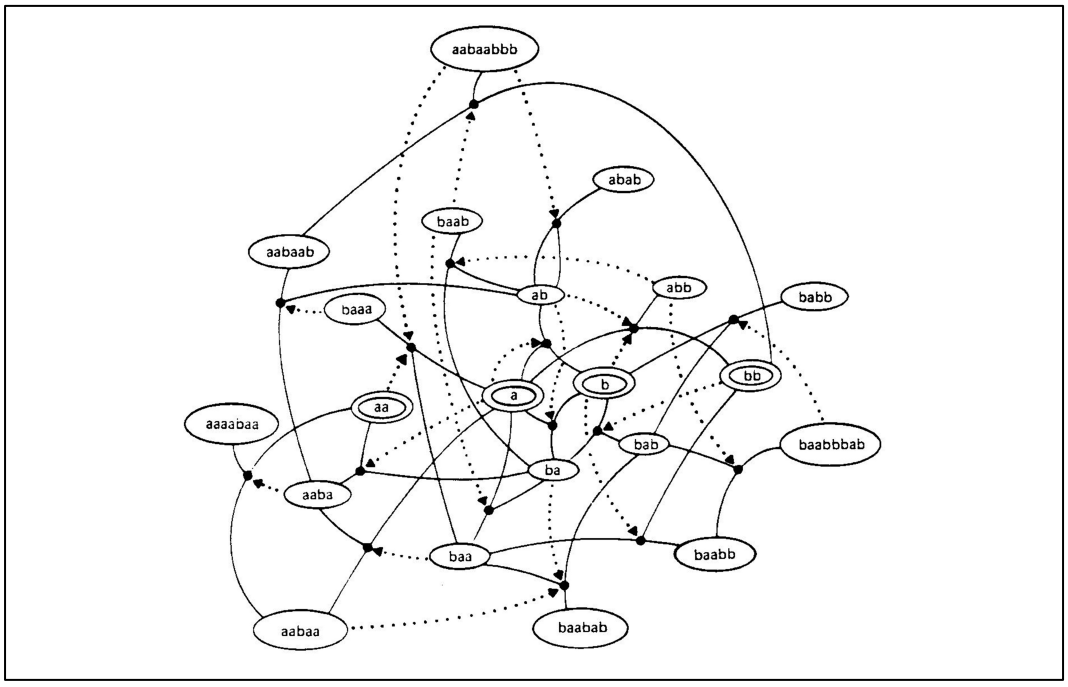


Figure 8.2: The relationships that comprise a collectively autocatalytic set. Small black circles represent ligation reactions whereby two smaller molecules are joined to form a larger third one. Dotted arrows represent the catalytic roles played by particular molecular species; each such dotted arrow connects a catalyst to the reaction it catalyzes. Each reaction is catalyzed by one member of the set while synthesizing another, such that the set as a whole is collectively autocatalytic even though the individual reactions may not be. Reprinted from Kauffman (2000).

Eigen's notion of a *hypercycle* (Eigen 1971; Eigen and Schuster 1977, 1979) describes another structure, more or less the same as Kauffman's sets in terms of their reflexively autocatalytic internal relationships. However, in contrast with an autocatalytic set, a hypercycle is both topologically more specific—consisting of an ordered cycle, rather than a network—and molecularly more specific, in that the components in Eigen's model are solely enzymatically-catalyzed RNA strands and the enzymes that they code for (see also the RNA-world hypothesis hinted at in Crick 1968; Orgel 1968; and Woese 1967; and stated most boldly in Gilbert 1986). Still, what is most important about the hypercycle is that the behaviors of its components collectively support one another's existence.

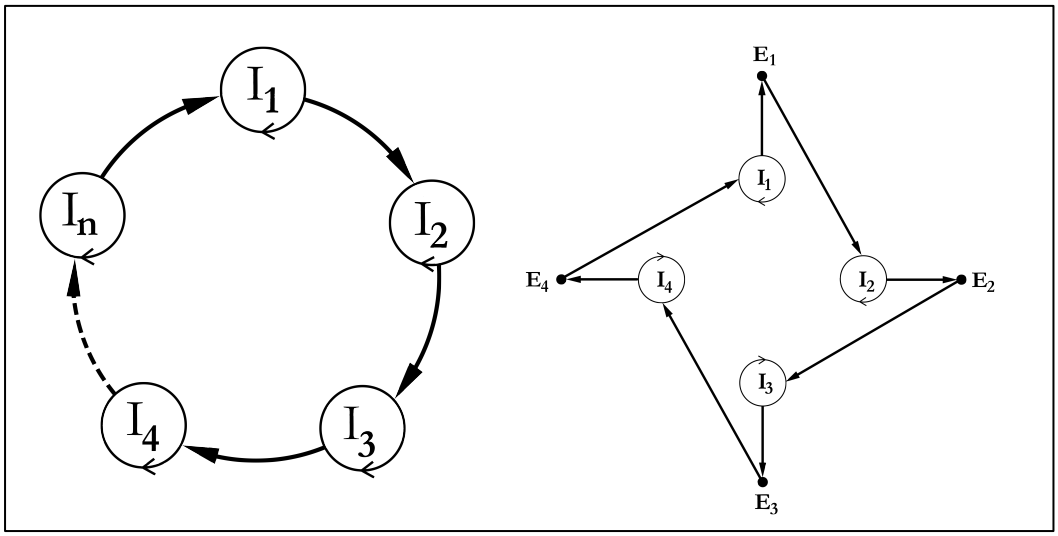


Figure 8.3: Two representations of the concept of a hypercycle. In both diagrams, the various encircled I_n signify autocatalysts—molecules that individually play a role in their own construction. The entirely circular arrows signify this level of catalysis. The remaining arrows also signify catalytic relationships, although these ones are not individually reflexive, but instead collude with one another to form larger collectively reflexive cycles. In the diagram on the left, the autocatalysts (I_1 . . . I_n) catalyze the synthesis of one another in a cycle; in the version on the right, the various I_n catalyze the construction of intermediary enzymes (shown as E_1 to E_4) which each then go on to catalyze the next autocatalytic process in the cycle. There is of course no reason why the enzyme-free and enzymatically-mediated modes could not also be mixed in a hybrid kind of hypercycle as well. Figure adapted from Eigen and Schuster (1977).

The most generic concept coming out of this faction of early-1970s theorists was Maturana and Varela's notion of *autopoiesis*—a coinage borrowing from the Greek roots for "self" and "production" (Maturana and Varela 1973/1980; Varela, Maturana, and Uribe 1974; see also Thompson 2007). Autopoiesis is described as:

A network of processes of production (transformation and destruction) of

> components, that produces the components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it [the autopoietic system] as a concrete unity in the space in which they (the components) exist (Maturana and Varela 1973/1980).

In short, what this means is that an autopoietic system is a set of processes that are self-producing. By using the term "processes of production", Maturana and Varela avoid committing to any particularly limiting type of causal behavior. Catalysis counts, but so do other cellular processes and even high-level organismic behaviors such as, say, making friends and sewing clothes, as long as the sum results of a group of connected processes are the continuous rebuilding of those processes themselves and the continued existence of the relationships between them. In archetypically Kantian fashion, an autopoietic structure is one that, through its various behaviors, attempts to ensure its own existence.

Autopoiesis and Identity

The central trouble with the theory of autopoiesis, as I see it, is that, even though Maturana and Varela recognize the importance of defining a notion of a self (or identity), the theory has yet to make that definition more than impressionistically. The problem of identity, recall, is that physical items may have countless minor variations, and there is no easy way to determine which variations count as being the same thing and which ones are sufficiently different as to be called something else. In the messy world of cells that are composed of many millions of atoms, nothing is ever quite the same; things can only ever be *very similar* at best. How could we ever know which processes that

have been created are the ones that did the creating? Furthermore, is that notion even coherent when charged with the responsibility of accounting for developmental change, such as that which occurs when a caterpillar metamorphoses into a butterfly? Certainly relatively few of the processes that make up the caterpillar are the same as those that compose the butterfly. And *Lepidopterans* aren't the only organisms with morphologically distinct life stages either; all organisms go through ontogenetic change in their life cycles, from infancy to adulthood. Even the individual cell—the paradigm autopoietic structure, the unit from which all other organisms are built—goes through vast changes during the routine course of binary fission wherein both its form and its processes vary widely from one moment to the next (see Figure 8.4).

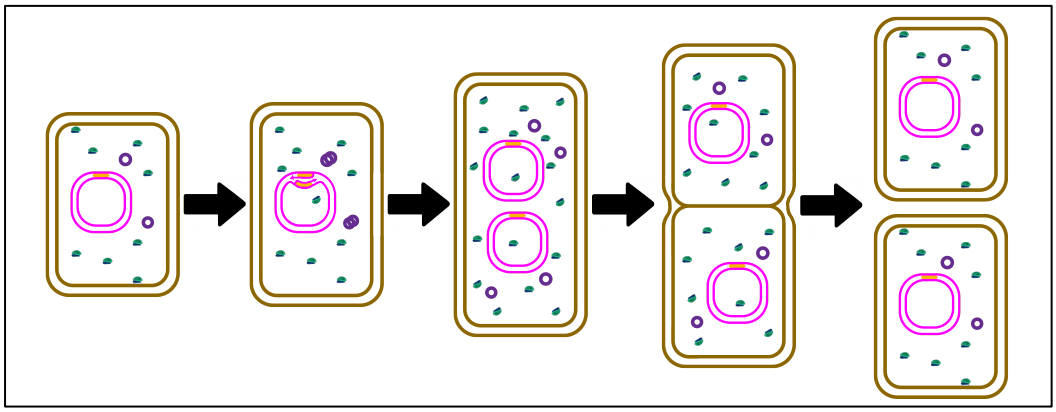


Figure 8.4: A schematic of the process of binary fission by which a prokaryotic bacterial cell grows and divides into two. The pink ring represents the bacterial DNA, the purple rings represent plasmids, and the small blue-green elements represent ribosomes. Many details have been left out for clarity.

Saying that a self is exactly the set of things that create themselves seems, at least at first, to be a superficially precise answer that does more to highlight the problem than to solve it. It makes a roughly outlined claim that the variations that we should count as being the same are whichever ones happen to work. The fundamental shortcoming of the theory of autopoiesis is that it depends

crucially on a notion of identity, which (so far) it doesn't offer. There is, however, a virtue in Maturana and Varela's theoretical generality, which gives many of their readers a keen sense that, somehow, they must be right: even if we don't know quite *which* things in an autopoietic system are the same over time, we do know, from observations of organisms and their life cycles, that *something* appears to be. I think that autopoiesis and the other models of autocausal persistence that we've just looked at can lead us to a theory of identity that may be able to account for, among other things, individual cellular continuity, binary fission, and the changing of a caterpillar into a butterfly.[302] We're going to follow that lead shortly now, but before we do, I want to point out the organizational features shared by the processes of autopoiesis and of replication.

Autopoiesis or Replication (Survival or Reproduction)

Autopoiesis is sometimes called "self-reproduction", and self-reproduction is typically contrasted with ordinary reproduction (also known as replication) in order to emphasize the two processes as distinct contributors to biological behavior.

To make the differences clear: autopoietic *self-reproduction* is the continuous renovation and remodeling of a structure to prevent it from falling apart, while replicative *reproduction* is the copying of a structure such that there may then be more than one simultaneous identical structure. The reparation of cellular damage is autopoiesis. Binary fission to produce two daughter cells is replication (see Figure 8.4 above).

While it is often more fruitful to focus on the differences between the two processes, there is also an important similarity. At the heart of things, each is an autocausal strategy of persistence,

[302] It is curious that, in the preface to Maturana and Varela's book (1973/1980), Stafford Beer was able to conclude, "The second reason why the concept of autopoiesis excites me so much is that it involves the destruction of teleology." That is exactly the opposite of the conclusion I would like to draw. I would say that autopoiesis is involved in the creation of teleology.

capable of preserving organizational information across time. In autopoiesis, a pattern that faces a background of braising follows an internal blueprint of a sort in order to repair damage, so that its identity (still not a precise term, here) continues to exist in the future. In replication, a pattern strives instead to make copies of its self, so that, statistically, even while some of those copies may suffer destruction at the hands of braising, others might continue to exist in the future. *Physically* the processes seem very different; but *informationally* both serve the same purpose of ensuring future existence by providing a level of informational redundancy. In both cases, the material in a structure may change but the information remains. Both processes start from a structure that houses redundant information, and both result in a structure that retains that redundant information, each using the only principle possible—the very redundancy within those structures—to protect its own organizational information against materially disorganizing damages. And because of this, structures that undergo either autopoiesis or replication or both all display the aspect of vitality, giving us the impression that these things are alive.

At this point, one might object that I seem to be overlooking an important difference between these cases: In autopoiesis the result is really "the same thing"—a persisting cell—while in replication the result is "just a copy". I see this concern as being rooted in unjustifiable prejudices about what constitutes an identity. If one were to observe both processes carefully, one would recognize that neither one turns out to be the persistence of a physical object. At some point in each process there is a complete turnover of physical material, and all that remains in the future is the same structure, the same organizational or informational content, the same pattern.[303] At any rate, this objection should dissolve when we have a coherent theory of identity.

[303] Over the course of an autopoietic cell's life, its every molecule is replaced thousands of times through continuous cycles of metabolic activity. In the process of replication by binary fission, autopoietic turnover continues unabated, and each copied daughter cell can only be made of roughly half the material of the original mother cell, with the other half being newly constructed parts . . . a process of change that accumulates geometrically with the passing of generations. In template replication, of course, the resultant copies are made entirely of new material.

Chapter IX

Identity

One string functions in two ways: first as program, and second as data. This is the secret of self-reproducing programs, and, as we shall see, of self-reproducing molecules.

—Douglas Hofstadter (1979, p. 499)

At least one generic model of identity can be found to support both autopoiesis and replication as forms of Kantian autocausal persistence. The proposal is based in part on the discrete mathematics of graph theory, in part on the dynamical-systems notion of a basin boundary, and in part on the notion alluded to in the epigraph above. Hofstadter's secret, "of self-reproducing programs, and . . . self-reproducing molecules", actually contains two of the secrets to any teleological reproduction or self-reproduction. The first secret is that one structure may serve two entirely different roles, once as "program" and once as "data". That is to say, in order to create anything—whether that is a copy of one's self or anything else—the creator needs to contain or have access to both a complete informational blueprint (the data) for what it is to create *and* the complete machinery (the program) that can follow that blueprint and do the creating.

The second secret, hinted at by the first, is that a set of patterns, arranged in the right way, is able to contain itself redundantly. Having two (or more) copies of the same organizational content simultaneously present allows one the luxury to lose any bit of information from one copy and to then potentially recover and reproduce the lost bit from the redundant information still available.[304]

[304] Computer scientists have, for some time now, understood the strategy for securing information by way of redundant storage. Not only do we have numerous procedures for backing up our data, but data-storage engineers have also embodied the notion of redundancy in a technology called Redundant Arrays of Independent Disks (RAID) used to

And so, by encoding the same content in two different ways, a set of interrelated structures can maintain, ensconced within their organization, the informational redundancy that is necessary to survive an environment of braising.

protect critical data systems. The general idea of RAID is that if our data is distributed across more than one physical disk, and if at least one of these disks is not additional data, but instead redundant data that is derived or calculated from the other disks in some reversible manner (commonly using a parity calculation), then whenever any single disk fails—either an original or the parity disk—the lost data can correctly be rebuilt from what remains on the other disks. In short, we can only rebuild lost information when we have informational redundancy somewhere in the system; otherwise, like a restoration artist, we would be left to guess.

A. The Subtlety of Sameness

If only by definition, it is impossible for two things, any *two things, to be exactly the same.*

—Robert French (1995, p. xv)

It should be obvious, I think, that the only possible way to account for the persistence of an identity over time is through some measure of equivalence that can group together a set of partially similar structures, allowing them to count as different versions of the same thing.

As we noticed in Chapter II, the majority of such equivalence classes that our human minds deal in are subjective; they are sets of structures that we can call "identifiable" because their equivalence exists in relation to some purpose for which the mind is making the identification (Hofstadter and Sander 2013; Lakoff 1987; Wittgenstein 1953). Examples are boundless, but include the numerous, largely similar (but certainly not identical) versions of Theseus' ship, or the many similar versions of a wedding ring as it weathers and accumulates years of scratches.[305]

Sameness is indeed subtle, and Wittgenstein's (1953) concern over the lack of essential categories in our world is of course concerning. But if every category were based only on Wittgensteinian family resemblance, then there simply would be no objective way for something to maintain its existence—nothing could ever remain the same, and there would be no true identities.

The seemingly pathological subjectivity of family resemblances is, however, not as dire as it seems. In certain cases there can be objective criteria by which to group the members of an

[305] The notion of subjective similarity is a deep topic that lies at the heart of cognition. It underlies the processes of perception, classification, and analogy-making that occur in our brains and, to some extent, also in artificial neural networks. For more detailed analyses of sameness and analogy, I refer the reader to Hofstadter and FARG (1995) and Hofstadter and Sander (2013), as well as to French's (1995) *The Subtlety of Sameness*, from which the current subsection takes its title.

equivalence class, in order to account for absolute identicality (and thus true identity). Although the epigraph above, from French, is true in part—there are no *physical* criteria according to which two things might be exactly the same—we may find two or more versions of a thing to be identical using *organizational* criteria. If each of the structures within an equivalence class is not required to be similar to the others but instead is required, in one manner or another, to have the capacity to produce or transform into one of the other members of the set, then they each have the transitive capacity to produce or transform into each other. There can thus be an objective measure of sameness across time. What remains the same amongst the members of such a set is not necessarily anything physical at any particular moment, but only the specification of what each member of the set has the causal disposition to produce—the identity is the entire set of potential future states, all of which can create one another.

Hofstadter's Secret

In developing his theory of *personal* identity[306], Hofstadter (1979, 2007) explored some apparently autocausal, self-referential patterns, a couple of which will be helpful to look at now to glean some intuitions about the types of things that are, in part, both cause and effect of themselves. Hofstadter's examples are not fully autocausal but, as we'll come to see, there is a useful analogy to be found between them and phenomena such as Kauffman's autocatalytic sets, along with the more general biological activities of autopoiesis and replication.

[306] Readers familiar with Hofstadter's work may discover some striking similarities between his notion and the one that I am developing, both of which are given similarly in terms of loops and cycles, self-reference, emergence, causation, and so on. It is far from accidental that these two theories, of topics that also share a central term in their titles, should converge upon such abstractly similar characterizations. Nonetheless, the two topics are quite distinct. Hofstadter's "strange loops" (2007) are phenomena that arise from a certain kind of self-referential structure within perceptual systems and so, in Hofstadter's sense, a tomato plant has no identity, a mosquito has nearly none, but a dog may have some, a two-year-old a bit more, and a healthy adult human will have something of the full-fledged sort. By contrast, the kind of identity I am working towards explaining is aligned less with awareness and perception and more with life and vitality; it inhabits mosquitoes and tomato plants equally as well as it does dogs, toddlers, and adult humans.

Hofstadter's first example draws upon Willard Van Orman Quine's *The Ways of Paradox* (1962/1976), an essay wherein Quine, explicating Gödel's (1931) proof of incompleteness, using an ingenious novel way of achieving self-reference, developed a crafty emendation to the ancient "liar's paradox" (also sometimes known as the Epimenides paradox).

The original paradox consists of a sentence such as "This sentence is false." Quine noticed that that seemingly self-referential statement might actually be found to be non-paradoxical, depending on what one takes to be the referent of the term "This sentence".[307] In order to remedy this, Quine devised a new sentence to do the same job better. I'll follow Hofstadter (2007), who labeled the new version of the paradox "Quine's Quip":

> "Yields a falsehood when appended to its own quotation" yields a falsehood when appended to its own quotation. (Quine 1962/1976)

If one were to attempt to check the truth-value of the assertion in Quine's Quip, one would take the quoted phrase and append it to its own quotation (as the quip suggests) and then check to see if that process yields a falsehood (as the quip also suggests). If this operation is successful—if the result that that process yields is indeed false—then the quip is true. But since performing the operation actually produces the quip itself, we run into paradox. As Quine himself puts it, the sentence "is true if and only if it is false" (Quine 1962/1976).

Hofstadter notes that the surprising and really powerful thing that Quine has done is something more than just to produce an interesting paradox: He has in fact done so by way of

[307] Quine (1962/1976) clarifies: "In an effort to clear up this antinomy it has been protested that the phrase 'This sentence', so used, refers to nothing. This is claimed on the ground that you cannot get rid of the phrase by supplying a sentence that is referred to. For what sentence does the phrase refer to? The sentence 'This sentence is false'. If, accordingly, we supplant the phrase 'This sentence' by a quotation of the sentence referred to, we get: '"This sentence is false" is false'. But the whole outside sentence here attributes falsity no longer to itself but merely to something other than itself, thereby engendering no paradox."

producing a thing that, in some sense, *is able to produce itself*.[308] Of course, without a person to follow its "instructions", Quine's Quip is not truly an independent replicator (and that fact will be important to us later), but it can be thought of, in a way, as a machine that is itself while simultaneously being the instructions to produce itself (Hofstadter 2007; see also Hofstadter 1979).

The other similar example that Hofstadter analyzed is also a thing that is able to produce itself, by way of the same special trick. In this case, the tricky type of thing is a bit of code—a string of characters—that, when run as a program on a computer, outputs the very same string of characters. Again in honor of Quine, Hofstadter dubbed this kind of program a "quine".

```
$s='echo "$s=".chr(39).$s.chr(39).chr(59).$s;';
echo "$s=".chr(39).$s.chr(39).chr(59).$s;
```

Figure 9.1: A quine written in the nonce computer-language qPHP. The first line of the code tells the computer to create in its memory a string variable (called **$s**) and assign it the value of the entire literal string of characters between the two "**'**"s. The second line of code tells the computer to **echo** (that is, to output) the value of that variable, essentially appended to its own quotation. What it echoes is a series of concatenated strings, beginning with a literal "**$s=**", followed by a single-quote mark [**chr(39)** means the 39th character in ASCII, which is just "**'**"], followed by the entire value of the string variable, **$s**, followed by another single-quote mark, followed by a semicolon [**chr(59)**], and ending once again with the value of **$s**. (The eagle-eyed reader who is familiar with computer code may have noticed that there is a line-break character after the first semicolon; that character is included here only to format the code so that it fits the page and is more easily understood.)

[308] This trick was first discovered and exploited by Kurt Gödel in his (1931) proof of the incompleteness of formal systems. But it is Hofstadter's analysis and expansion of Quine's version of Gödel's idea that will be useful to our current project. The interested reader should explore Gödel (1931), Hofstadter (1979, 2007), and Quine (1962) as well as von Neumann (1966) and Smullyan (1961).

The listing in Figure 9.1 is a sample quine, written in a language we can call qPHP, a purpose-built derivative of the recently popular scripting language PHP.[309] I chose this particular quine as an example, because it visibly operates in quite the same way as Quine's Quip. It has two halves that look almost exactly the same, despite their playing different roles in the act of replication (see also Hofstadter 1979).[310] One can follow the figure caption above to see in detail how the quine works, but in sum, the entirety of the program pretty much says:

> 'Yields a quine when appended to its own quotation' yields a quine when appended to its own quotation.

As Hofstadter explains it, the secret—the special trick that allows both Quine's Quip and these software quines to do what they do—is the fact that both kinds of structures play dual roles, serving at the same time as both data and program, both blueprint and machine. Quine's Quip and

---

[309] A few notes on syntax, for programmers: qPHP is like PHP in that it uses the convention of naming and declaring variables by prefixing their names with a dollar sign, so **$s** serves as a string variable here. Also, the "." is used as the concatenation operator, meaning to join the strings before and after it, end to end, into a single string. The two main differences between PHP and qPHP are the following: (i) in qPHP, variable interpolation has been turned off. This means that a variable name found inside a string literal will only be interpreted literally, instead of being converted to the value of that variable as it is in PHP. And (ii) in PHP, a programmer needs to tell the interpreter that there is code to interpret by enclosing the code in a block that begins with "**<?php** " and ends with " **?>**"; in qPHP everything in a file is expected to be code and so no enclosure is necessary. The "same" quine written in PHP is the following relatively longer block of code; curious readers who would like to work out for themselves how it operates would be well-served to consult an ASCII character-set encoding table.

```
<?php $str='echo chr(60).chr(63).chr(112).chr(104).chr(112).chr(32).chr(36).
"str=".chr(39).$str.chr(39).chr(59).$str.chr(32).chr(63).chr(62);';
echo chr(60).chr(63).chr(112).chr(104).chr(112).chr(32).chr(36).
"str=".chr(39).$str.chr(39).chr(59).$str.chr(32).chr(63).chr(62); ?>
```

[310] Hofstadter's own example quine (1979), written in a language he called BlooP, is also aesthetically constructed of an obviously reduplicated block of code wrapped in minimal syntactical punctuation:

```
DEFINE PROCEDURE "ENIUQ" [TEMPLATE]: PRINT [TEMPLATE, LEFT-BRACKET,
QUOTE-MARK, TEMPLATE, QUOTE-MARK, RIGHT-BRACKET, PERIOD].
ENIUQ
['DEFINE PROCEDURE "ENIUQ" [TEMPLATE]: PRINT [TEMPLATE, LEFT-BRACKET,
QUOTE-MARK, TEMPLATE, QUOTE-MARK, RIGHT-BRACKET, PERIOD].
ENIUQ'].
```

An English translation of this quine that respects the syntactical order of Hofstadter's example might look like the following: A quine is produced when one prepends to its own quotation the phrase "a quine is produced when one prepends to its own quotation the phrase".

our qPHP quine both contain within themselves the thing that needs to be written twice and, right next door, the instructions for how to write it twice, in order that, together, they might be able to produce a thing that is written twice.

Multiquines

In the cases we've been considering, the blueprint and machinery for replication stand side-by-side in a neatly reduplicated form. Their information is explicitly present twice, once in each half. But not all replicators require this explicit reduplication, and as we explore more, we'll find that, in most replicators, things are not so neatly divided. Nonetheless, the reduplication in the Hofstadter–Quine examples is curious, and it hints at the idea that some form of redundant information may be a requirement for replicators.

We can take a different perspective on the role that redundant information plays in replicators by looking at cases in which a thing does not directly create itself. Consider what are now typically called multiquines—code for computer programs that write code for other computer programs (potentially, but not necessarily, in other languages), one of which eventually will write the code for the original program again.[311] Figure 9.2 contains listings of programs P and Q that are able to write one another, both in the javascript language. Each program certainly contains an informational specification for itself, just by being there; it is itself. And it should be intuitively obvious that each program also, in some fashion, contains the organizational information—both blueprint and machinery—required to produce the other. But what that fact implies is that, in some

[311] An engineer named Yusuke Endoh produced the most virtuosic multiquine that I have yet run across. Endoh included ninety-nine programs as embedded data in each functioning program, for a total of one hundred programs altogether, each of which writes the next program in a circular series. Every version of Endoh's multiquine is a different structure—a physically and logically different string of bytes, in a different language—with the capacity to produce each of the other programs and thereby to, eventually, produce itself (Endoh 2014, 2015).

fashion, each program must additionally contain the organizational information that is key to its own creation.

PROGRAM P

```
function p() { console.log("function q() { var b = " +
String.fromCharCode(39) + p.toString() +
p.toString().substr(8,4) + ";" + String.fromCharCode(39)
+ "; console.log(b); } q();");} p();
```

PROGRAM Q

```
function q() { var b = 'function p() {
console.log("function q() { var b = " +
String.fromCharCode(39) + p.toString() +
p.toString().substr(8,4) + ";" + String.fromCharCode(39)
+ "; console.log(b); } q();");} p();'; console.log(b); }
q();
```

Figure 9.2: A two-part multiquine. Programs P and Q, both written in javascript, are designed such that they write one another's source code. While both programs use string manipulations to write code out to the console, their particular methods of building one another differ slightly. As one can see, a large part of these two programs is the same, indicating that they share much of the same organizational information. In each of the two programs, the portions in bold typeface indicate the verbatim contents of Program P. The red, blue, green, and purple characters in each program represent the corresponding, encoded contents of Program Q.

If we look at the programs in Figure 9.2, we can easily see that the entirety of P is located inside Q, as a literal string. We can call this *the P that's in Q*. I've highlighted the P that's in Q by putting the verbatim segment of each program in bold type.

But in order for each program to be able to write the other, the entirety of Q must also reside, in some fashion, within P. In fact it does, and it is not too difficult to see how. The blueprint for Q can be found in P not literally, but as a set of encoded string operations that can construct Q. Observe: The beginning of Q is there in P, literally: "**`function q() { var b =`** ". And so is the ending: "**`; console.log(b); } q();`**". I've made these corresponding sections within each program red so that they can be easily compared. Then, the two single-quote marks of Q are encoded in P as "**`String.fromCharCode(39)`**", (which is javascript code for printing ASCII character number thirty-nine) and there is one functional syntactical semicolon, "**`;`**", in Q that is encoded as a string literal in P. All of these items have been colored blue in each program. The remainder of Q just *is* P, as we've already noted, but the way that that part of Q is encoded as a blueprint within P is a bit trickier than the literal way that P is encoded in Q. In this case, the bulk of that part of Q is encoded as "**`p.toString()`**", which is a shorthand in javascript that asks the interpreter to get the string that corresponds to the code for function p. I've made the corresponding parts in each of the two programs green. And the small remainder of Q not yet accounted for is encoded as "**`p.toString().substr(8,4)`**". That part of P and the small piece of Q that it encodes—a four-character substring of the code for the function p, beginning from the eighth character—are both in purple. Altogether, just as the bold section of P is the P that's in Q, the colorful section of P represents *the Q that's in P*.

Now that we've seen how P is inside Q and how, at the same time, Q is inside P, it might strike us that this means there should, in some fashion, be a Q deeper within the P that's in Q as well as a P deeper within the Q that's in P. Indeed, this is the case. For instance, if we look within

the P that's in Q—that is, if we look at the literal string in bold type inside Q—we can find a representation of the code that corresponds to Q, just as we described above for the Q that is inside P itself. I've underlined the relevant piece. This is *the Q that's in P that's in Q*. And if we look again at the Q that's in P—the series of colored string-manipulating code statements within P—we find within them also a further version of P, encoded as the following subset of the commands: **`p.toString() + p.toString().substr(8,4) + ";"`**. That is *the P that's in Q that's in P*.

Altogether, each program not only (i) is itself, but also (ii) encodes the immediate or proximal potential for its partner or successor, and (iii) is also self-redundant, by encoding the distal potential to make a structurally identical copy of itself again. Both programs indirectly contain versions of themselves. They have informational redundancy (even without containing the obvious reduplication found in quines and Quine's Quip). They have the power to cause their own existence by way of causing one another's existence.[312]

Multiquines, Replication, and Autopoiesis

Multiquines make for a rich example within which we can find versions of both replication and autopoiesis. Consider a few different possibilities. First, imagine we begin with a single copy of a single version of a multiquine (say, program P), and when that program is run, it produces the next version (Q), after which the first program is cleared from the computer's memory. Under this constraint, no matter how many versions there are in the multiquine, we have a system that is just as

[312] A multiquine is a good example for beginning to explore the concepts of actual and potential organization but, ultimately, these examples suffer from one of the problems that also afflict quines and Quine's Quip. None of these things are perfect examples of the idea we are using them to represent because they aren't truly standalone machines that can follow their own blueprints to produce themselves. They are all *instruction sets* that require rather complex auxiliary machinery—people and possibly computers—in order to operate. They are like prions or virions without hosts, or cellular DNA in the absence of ribosomes and mitochondria. They are simply not autocausal sets of their own.

susceptible to braising as any other program because, no matter which version exists at the moment, if it becomes corrupted—if almost any character within it were to become displaced within the computer's memory—it will no longer be able to write the next version (and so our only copy of all potential versions will be lost at once). All these individual versions of a multiquine contain the organizational potential for all the other versions but, arranged and allowed to operate in this way, they are still fragile, because their redundancy is not distributed. The organizational potential for all versions is located in one place—in engineer's terms, a single point of failure. We can represent this state of affairs as in Figure 9.3.

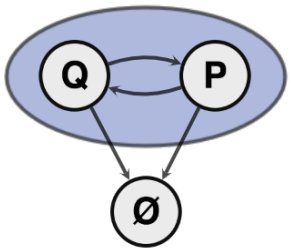


Figure 9.3: A representation of programs P and Q, which are able to sacrificially write one another's code. The arrow from Q to P represents the process by which Q writes P and is then, itself, destroyed. The arrow from P to Q represents the same process in the other direction. And the arrows from each of P and Q to Ø represent the possibility of either of these two programs decaying due to random outside events (we can think of Ø as being the "null state" where neither program exists). We will develop the notation that you see here (which I call an "organizational graph") in detail as the chapter continues.

Alternatively, we can imagine a possibility wherein each version of the multiquine is allowed to run and to write its own output as a new file, but only one copy of each version is allowed to reside on the disk at any time. What we get is something akin to an autocatalytic (or autopoietic) set.

Although there is only one copy of the entire set of programs, that one copy contains a lot of redundant data, since each version within it contains the blueprints for all the others. To be precise, there are as many copies of the information for each version as there are versions. This distributed redundancy protects all the versions against random and irreversibly destructive events, because any one of those versions contains the potential to recreate all the others.

Say, for instance, that program P is able to write program Q, while Q is able to write P. If for whatever reason, the version of P in storage becomes corrupted, it can be reproduced by the future operation of Q . . . or vice versa. The various programs that make up the multiquine can be thought of as various parts that have the ability to repair or replace one another.

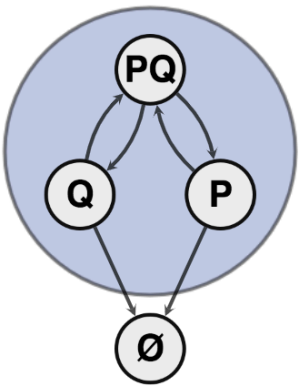


Figure 9.4: A representation of the same two programs, but now arranged in a scenario in which they both may coexist in the computer's memory. When P creates Q it can also stick around, resulting in a state with both programs (PQ), and vice versa.

Lastly now, we can imagine a scenario in which each version of the multiquine is retained after running and is also allowed to write its output as a new copy. Now we have an autopoietic replicator. In this case, as long as the rate of externally produced damage is slower than the rate at

which the programs run, the set of all versions is not only able to maintain itself (despite occasional damage to some versions) but it is also able to fill up the disk with endless copies of itself, providing a further level of protection against braising—if the new versions are written to new sectors of the disk or to new disks altogether, then a sector failure or a disk failure (instead of a bit failure) can be tolerated, and the autopoietic replicator will continue to exist, or may even proliferate.

As we continue, the proposal I'll be making is that autocatalytic sets, hypercycles, chemotons, and other autopoietic entities share these informational properties of the multiquine—their multiple parts contain the blueprints for one another (via their capacities to construct one another) and thereby contain the blueprints for themselves. The difference that makes physical autopoietic entities even more interesting than multiquines is that they are not just instruction sets—they themselves also contain the machinery necessary to follow those instructions. The various versions of any of these structures have the power to cause their own existence by way of causing one another's existence. To make that claim more meaningful, we can talk a bit more now about what remains the same amidst all this change.

Organizational Identicality

From the example of the multiquine, we begin to get the sense that what it means to have an identity is not necessarily to remain *physically* identical as time passes, but to remain *organizationally* identical . . . to possess, in every possible version of a changing thing, the deep, distal blueprints—the organizational capacity—necessary to eventually create any of the other versions.

It will help us to understand organizational identicality if we make a distinction between what I'll call *actual* and *potential* organization. The *actual* organization in a state is the physical, structural organization of the patterns that actually exist at that moment. It is some kind of

specification of what is really there, and it is what differs from state to state.[313] The *potential* organization in a state refers to the ability of that state's actual organization to causally contribute to the organizational content of future states. In an identity, it is this organizational potential that remains the same between states.

The set of programs that constitute a multiquine are quite distinct if you compare them side-by-side, byte-by-byte. They are conceivably written in different languages, translated by different interpreters or compilers, and they result in differing series of operations to be performed by the same processor. But, as we know from their performance, they are all inter-convertible. Each version directly specifies its own actual organization, directly encodes the potential organization for the next program in the sequence, and thus indirectly encodes the distal potential organization for the rest of the programs in the sequence including, ultimately, itself. What this means is that each one of them contains the potential for creating all of them, and so *the sum of the potential organization in each one is precisely equivalent*. The versions of a multiquine are not physically, but *organizationally* identical.

Blueprints and Machinery

We are soon going to apply the notion of organizational identicality to examples that, while naturally abstract, will also be able to map onto chemical systems. It will be useful to preface that discussion by looking at how actual and potential organization might arise in physical and chemical systems.

[313] One day someone may discover an information-theoretic (*i.e.* bit-string) formula for encoding organizational patterns as bit-strings. If so, I would conjecture that what it means for different states to be organizationally identical is for each of the states to be equivalently compressible to the same minimal bit-string. At present, I haven't the faintest idea how that would be done.

By now, physicists have well documented the nature of matter as *stuff* that can *interact*—as patterns that can cause changes (to one another). Matter itself is at the same time both object and motion, both noun and verb, both data and program; it is a thing unto itself and simultaneously a capacity to change other things; a thing that can move and a thing that can be moved. And so here we can reframe the dynamical account of patterns from Chapter II in terms of actual and potential organization: Each pattern of physical matter in the world—whether it is a particle, a small molecule, or a highly complicated aggregate—embodies both its own actual organizational structure and, at the same time, some specific causal capacities—some potential organization. Just by being a pattern of matter, a thing is inherently endowed with the dynamical, causal blueprints that are able to create a certain set of other patterns (given an environment with the right material and energetic components).

We can make this more specific with an example from chemistry: A volume that contains a polycrystalline lump of platinum (Pt) along with some molecules of gaseous ethene (C_2H_4), and some molecules of gaseous hydrogen (H_2) has an actual organization that can be described in just those terms. But latent in that mixture also lies the potential organization for ethane (C_2H_6), a compound formed by the catalytic action of the metal upon the two gases. There is no ethane in the mixture at first, but nonetheless there are both the blueprint and the machinery necessary to create it—the mixture might be said to be "pregnant" with the organizational potential for ethane. And sure enough, in a range of energetic environments, the starting state described above will usually evolve into a new state that contains ethane (along with the platinum catalyst, which itself goes unchanged during the hydrogenation of ethene that it encourages).

This example is *not* autocausal in the way a multiquine might be, because the later state (ethane plus catalyst) is not likely to produce the former (hydrogen, ethene, catalyst); however, the example does demonstrate the difference between the actual and the potential organization of states.

In addition, it allows us to imagine arranging chemical systems into Kauffman-style autocatalytic sets that *are* like multiquines, in which the potential of each chemical state contains the distal blueprints for the remainder of the states, including itself.

B. A Mathematical Form of Identicality

Now there are selves. There was a time, thousands (or millions, or billions) of years ago, when there were none—at least none on this planet. So there has to be—as a matter of logic—a true story to be told about how there came to be creatures with selves.

—Daniel Dennett (1989)

But to think [of an item, such as a glass, as having a permanent nature] is to fall into the trap of Plato's "objectivist" vision, according to which objects have one and only one true identity.

—Douglas Hofstadter and Emmanuel Sander (2013)

Although the notion of organizational identicality that we are about to develop is mathematical, it is better thought of as a kind of logic than as a formula or model for a particular structure. It is an exploratory version of a tool for reasoning about the causal interactions that allow various kinds of patterns to create one another, and for categorizing the sources of orderliness in those patterns. The design of the tool can probably be improved upon; but this first version seems to be *worth improving upon* because it does us the service of bringing together the conceptual pieces that can help us understand identity and value (and all that comes with them). Since my notion of identicality follows from the discrete mathematics of graph theory, I will begin with a brief introduction to those abstractions.

A *graph* is a mathematical representation of *items* and their *relationships*. The items in the graphs we'll analyze will be symbolic representations of the *actual organizations* of patterns; the

relationships will be the capacities of those items to transform into one another by way of physical causation. But embedded in those representations, in a way that I'll soon describe, will be a further representation of the *potential organization* of those patterns. Typically, a graph will consist of *nodes* (or vertices; drawn as circles) that represent the items, and *connections* (or edges; drawn as lines or arrows) that represent the relationships, but alternative representations (for instance, matrices) are sometimes used to denote the same information.

A graph can be as simple as a single node—a symbol on a page, usually circled—that represents whatever the person who drew the graph has decided for it to represent. Two circles connected by a line represent different items that are somehow related to one another. For instance, one circle might represent Madrid and another Paris, and the line between them might have been chosen to represent the idea "has a direct flight route". Another node, perhaps representing Honolulu, may have no edges connected to it, thus representing the idea that there are no direct flights between it and either Paris or Madrid.

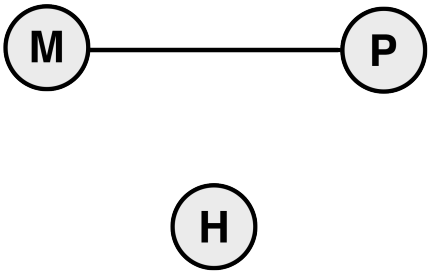


Figure 9.5: A graph representing the existence of a direct flight route between Madrid and Paris, as well as the lack of direct flights between either of these cities and Honolulu.

A *directed* graph is one in which the connections are arrows, rather than lines, symbolizing the potentially one-way nature of a relationship. Thus, however likely or unlikely it might be for airlines to do so, there could, for instance, be direct flights from Madrid to Tripoli, from Tripoli to Paris,

and from Paris to Madrid, but no flights going in the other directions, and we can symbolize this state of affairs by drawing a set of nodes connected by arrows. The relationships represented by the arrows in a directed graph no longer represent "has a direct flight route", but instead represent the more specific idea "has a direct one-way flight" (in the direction of the arrow).

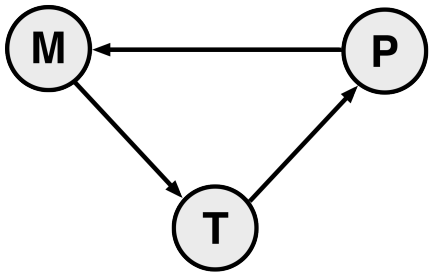


Figure 9.6: An example of a directed graph, representing the flights available from Madrid to Tripoli, Tripoli to Paris, and Paris to Madrid. Obviously, given this schema, if a Madrilenian would like to visit a Tripolitan friend, they will have to stop over in Paris on their way home.

A full description of the world's major cities and the available air-travel routes between them would of course have many hundreds of nodes and many thousands of connections. In general, a graph may have as many nodes and connections as are necessary to represent the structure one is attempting to describe.

One feature of directed graphs that is important to the notion of identity we're going to develop here is that there is the possibility of a *cycle* of directed edges that allows one to pass through some number of nodes within the cycle, eventually returning to a previous state. The most minimal cycle consists of a single node with a self-connection (look ahead to Figure 9.8, for an example). Another example appears in Figure 9.6, where the three nodes together form a cycle.

Cycles are of course interesting because, in going round and round, one may come back to where one started, thus offering a mildly tantalizing notion of sameness; but cycles get to be even

more interesting when taken to the next level: Graph theorists have chosen the name *strongly connected component* (SCC) to refer to a group of nodes within a graph having the property that every node can be reached from every other node through various interconnected and partially overlapping cycles. One feature of SCCs that differentiates them from individual cycles is that, while the nodes within an SCC are all mutually reachable, none of them have a reciprocal connection—direct or indirect—with any node in any other SCC of the same graph. That is to say, an SCC is an *isolated* set of connected cycles; both entrance into and exit from an SCC are always irreversible (one-way) transitions. This feature of directed graphs produces a kind of *natural, mathematical boundary* around members of the SCC that marks those members as somehow being *the same*. At the same time, the boundary also marks the rest of the nodes in the graph as somehow being *different*.

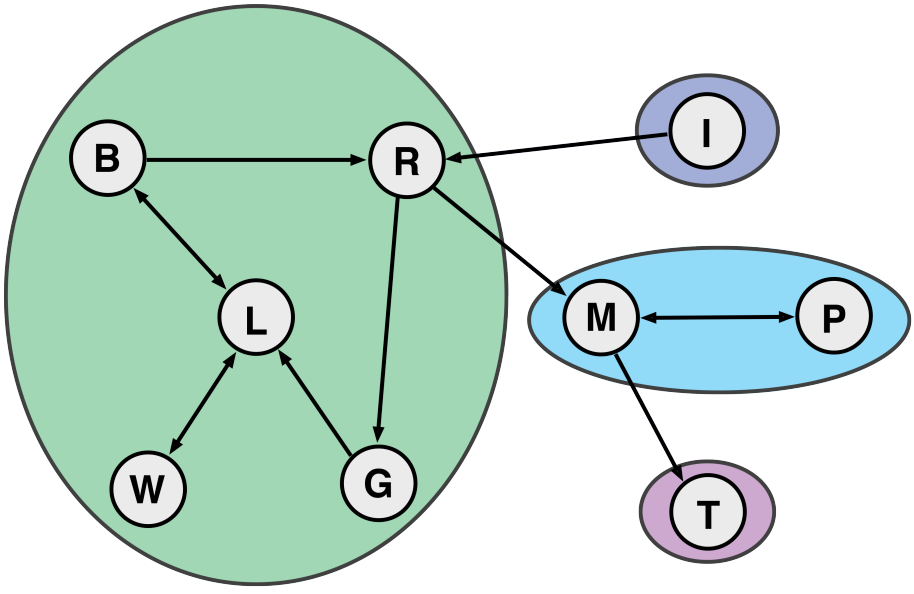


Figure 9.7: A directed graph with four strongly connected components (SCCs) highlighted. As is apparent in the example, one can find a path from any node within a particular SCC to any other node of the same SCC (and back), yet only one-way paths exist between nodes in separate SCCs.

A further result of the boundaries of strongly connected components is the fact that any directed graph can be partitioned into a set of mutually exclusive strongly connected components. That mutually exclusive set of SCCs, and the one-way connections that remain between them, serves as a directed graph of its own—called the *condensation* of the original graph. Of course the condensation of a graph is always acyclic (otherwise the SCCs in it would be joined, not distinct). In figure 9.7, the condensation of the graph has four nodes (the four SCCs of the full graph) that can be arranged in a serial fashion along the lines of their connectivity. The source node, I, of the condensation transitions to node B–G–L–R–W, which then transitions to node M–P, which at last transitions to the sink node T (this kind of serial pattern is not typical; it just happens to hold in this case).

In the current work, our analysis of identity will be given in terms of the SCCs that occur within a special kind of directed graph whose nodes and connections represent both organization and time. These graphs fit into a class of models called *Markov processes*, which are generally meant to model time. The Markov processes we'll look at can be used to represent the changes that might occur to a set of organizational structures over time, and the dividend paid for taking that point of view is that we can also look within these models for SCCs that represent *patterns of potential organization that don't change over time*. That is to say, in graphs where the nodes represent actual organization, an SCC represents a set of nodes all of which have the same organizational potential to create one another—they are like the multiple versions of a multiquine, each P containing some *proximal* potential (the Q that is in P) as well as some *distal* potential that includes itself (the P that is in Q that is in P). It is that persistence of organizational potential that we'll consider to be the abstract heart of organizational identicality.

Time

So let's look now at how these Markov-process graphs can help introduce time into our analyses. In a Markov process, each of the nodes stands for a description of a state of the system at a particular moment. The directed connections from node to node signify the potential transitions between organizational states as time passes.

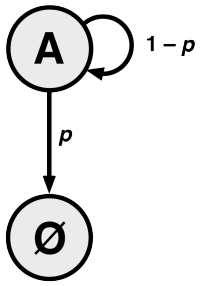


Figure 9.8: A directed graph representing the Markov process in which state A transitions to state Ø with probability *p*, and also transitions to itself with probability 1 – *p*. Whenever the sum of the probabilities for all the transitions outbound from a node is less than 1, we assume that the node has a self-connection with a probability that brings the total to 1. In the case in which there are no outbound transitions from a node (such as node Ø in this model), the node is called an absorbing state and the implicit self-connection has a transition probability of 1.

I've drawn the very basic Markov model shown in Figure 9.8 as an example to help us understand this way of representing time. In such a model, we begin in one of the nodes at time t_0 (we assume whichever node we begin in represents some initial conditions of our world) and, with each discrete tick of a clock (t_1, t_2, t_3, . . .), a probabilistic decision is made as to which of the transitions to follow in order to reach the next state of the system. In other words, as time goes by,

the world either remains in the same state (at the same node) or it changes to an allowable new state by following an arrow.

For instance, one could consider the state A to represent the fact that a coconut tree remains upright and rooted during a typhoon. The model in the figure claims that, at any moment in time, there is some small probability *p* that the tree may get blown over into state Ø. The remainder of the time (probability 1 – *p*) the tree will remain standing, and so a self-connection (a curved arrow going from state A to itself) represents this maintenance of the status quo. There is no arrow from state Ø to state A, and this represents the fact that the effects of the wind (which is the causal factor lying behind our transitions) cannot re-root a fallen tree (*i.e.,* the probability of such a transition is 0).

The transitions taken are stochastically chosen, according to the transition probabilities of all the outbound transitions from the current node. For now, we will look at discrete-time Markov models that are time-homogeneous—that is to say, there is a finite, fixed set of states, and as time ticks by in discrete fixed units (called "ticks"), the models transition from one state to another.[314]

A Discrete Representation of our World

Throughout much of the rest of the chapter, we will be attempting to determine whether various combinations of physical patterns can be considered to be organizationally identical. In order to do so, we will use a system of symbols, just like the letters in the graphs we've been looking at so far, to represent each of those patterns and the causal relationships that may exist between them. Before settling into our use of that symbolic system, however, we should convince ourselves

[314] These assumptions certainly oversimplify most real-world systems; however, our topic already has many pieces to be understood and, while Markov processes can also be described in terms of continuous-time dynamics, the discrete-time simplification will be enough to jumpstart our explorations and to help us forge many of the intuitions we need regarding identity, value, and goal-directedness.

that it might be reasonable to make the leap from the messy, continuous dynamics of our world to the use of those discrete symbols.

I'm going to try to make one plausible suggestion as to how to make this mapping. However, I don't want the remainder of the work to stand or fall based on this one suggestion, so first of all, I would like to point out that there may be other mappings of the continuous onto the discrete that could do the job just as well or better, and if a more convincing mapping can be found by scientists more well-versed in physics than I, then so be it. The major notions of identity and value that result from the discrete mathematical analyses that follow can be judged independently from whatever mapping is used to connect those abstractions to the real world. With that said, let's look now at the mapping I propose.

We commonly talk about molecules as if they are rigid, individual patterns, the way we envision them, for instance, when using the ball-and-stick models of introductory chemistry classes. In reality, however, those ball-and-stick models are simplified representations that stand in as prototypes for large categories of very similar structures. Every molecule is a constantly vibrating, twisting combination of atoms. And each atom has an ever-changing, vibrating nucleus, surrounded by an endlessly swirling "cloud" of its electrons.[315] The thing that justifies the use of those ball-and-stick models, as well as the symbolic chemistry that we often do on paper, is that generally those many variations all tend to behave in the same way. For whatever reason, a hydrogen atom reliably behaves as a hydrogen atom, and a carbon atom as a carbon atom, and so on. The point I will try to make here is much the same as that commonplace observation, but I will try to spell it out in a little more detail, and also tie it in a little more closely to our graph-theoretic analyses.

[315] Despite how tempting it may be to try to involve quantum mechanics in this analysis I am going to leave aside such quantum issues as the discrete energy levels of electrons as they orbit a nucleus. There may be something of value there but personally, I don't yet see how one might generalize those concepts to all states of organization.

If each of our symbols stands for some prototypical atom or molecule, and if every such structure is really an uncountable collection of continuously varying microstates made of that structure's parts, then, when we consider, say, two of our symbols to combine into a third, we would like that capacity to be independent of the particular microstates of those patterns that existed at the moment of combination. We would like it if any version would do as well as any other. The chemist's success with naming atoms and molecules—and predicting their behaviors categorically—already makes this view highly appealing, but we can put together a few pieces that may help us better understand it at a lower level.

The first piece is that the nature of physics itself naturally discretizes the world into two categories of states—those that are *bound* and those that are *unbound*. Whether we are talking about nuclear (say, proton-to-proton) attraction, electromagnetic (electron-to-proton; atom-to-atom) attraction, or gravitational (massive body) attraction, any two patterns have what is called a potential energy well into which their state of relation might fall. Each of the infinite possible states of relation between two patterns (in terms of, say, their masses, charges, distance, and relative velocities), either falls into that potential energy well or remains outside of it, thereby collapsing the continuous state space into our two discrete categories: bound and unbound. And so, in the absence of outside forces, two atoms, for instance, are either certain to ultimately move apart (they are unbound) or certain to come into one another's orbit (they are bound).

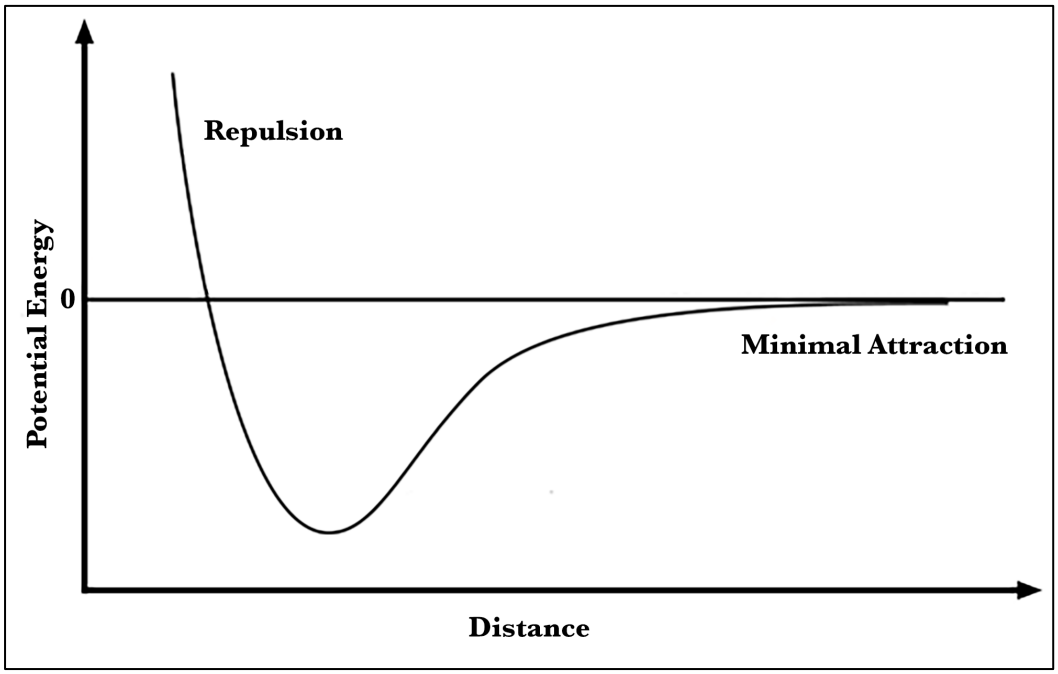


Figure 9.9: A (coordinate) graph of a potential energy well, showing the way potential energy between two particles varies with distance. Ideally—in the absence of outside forces—the particles will be drawn toward the bottom of the well, to that particular distance that corresponds to the minimal potential energy. Whether or not the particles are actually bound to come together there is determined also by whether the kinetic energy—the current motion—of the particles drives them towards or away from one another more weakly or more powerfully than the forces that define the shape of this well. If the kinetic energy is great enough, the particles are unbound, and if it is weak enough, they are bound.

When the state of relation between two particles does fall into that potential energy well, the particles become quickly constrained to a small subset of their total possible states of relation. Without additional energy, they won't be able to get over the walls of the well, and so the distance between them becomes limited to the tight range specified by the width of the well. Because of this, the combined shape they produce, while still vibrating somewhat, does not vary too significantly.

We might be able to see this in another way by putting the same idea into the language of dynamical systems theory. In terms of dynamics, we can say that the potential energy well produces

an *attractor* (in some coordinate space, perhaps that of distance and velocity between the particles or the atomic nuclei). Around any attractor is what's called a *basin of attraction*—the region of that coordinate space wherein all flow heads towards the attractor. And around any basin of attraction is what's called a *basin boundary*—a line of demarcation within which lies the basin and beyond which flow is exempted from reaching the attractor. This basin boundary serves as the sharp line that distinguishes a bound state from an unbound state. In terms of the kinetic energy, the potential energy (in the forces of attraction and repulsion), and the distance between our particles, the entire physical relation between the particles either lies at a point within the basin boundary, where the particles can be considered bound, or it lies beyond the basin boundary where the particles can be considered to be unbound. I suggest that the reason that particles, atoms, and most small molecules appear to be natural kinds—or at least, to be much closer to being natural kinds than larger aggregations of matter—lies in the way the dynamics that hold them together draws these sharp lines, determining which states of relation count as being the same and which count as being different. Existence as a particular pattern of particles can be thought of as a discrete, binary affair. A pattern of this sort either exists or it does not.

The idea of discrete causal relations that we are going to depend on for our upcoming analyses is that some patterns, perhaps X and Y, may have some likelihood, p, of creating some other pattern, say Z. Based on the foregoing, one thing we can account for in this picture is the idea that in order for Z to form, creating any state in the basin of attraction for Z seems to be as good as creating any other state in the same basin, because the many states will quickly converge to the same attractor. However, we still need to account for the opposite idea—the notion that any versions of X and of Y will serve as well as any other in being able to form Z. The many vibrational variations of a lump of platinum, for instance, can all catalyze the hydrogenation of ethene into ethane, and the

many vibrational variations of a hydrogen atom can equivalently bind with any of the variations of a fluorine atom to form hydrogen fluoride.

We can augment the notions so far with a probabilistic argument. As we've determined, because of the forces that bind their parts together, the various states we are considering to be equivalent are all fairly similar in shape and able to perform the same causal jobs in intermolecular interactions. However, if for some reason some of those variations are not able to perform those jobs, they nonetheless transform from one to another so quickly that any one of the variants that fails to perform a job will soon be replaced by one of the variants that can perform the job. One might be reminded, here, of the way the lock-picking technique called "raking" is able to quickly shuffle through many variations of pin heights within the tumbler and, sometimes in mere moments, find and effectively imitate the shape of the key. The rake doesn't need to be shaped just like the key; it just has to cause quick enough random change (centered on the average shape of a key) such that the shape of the actual key will soon come to exist, for at least a moment, and push the system over the basin boundary.

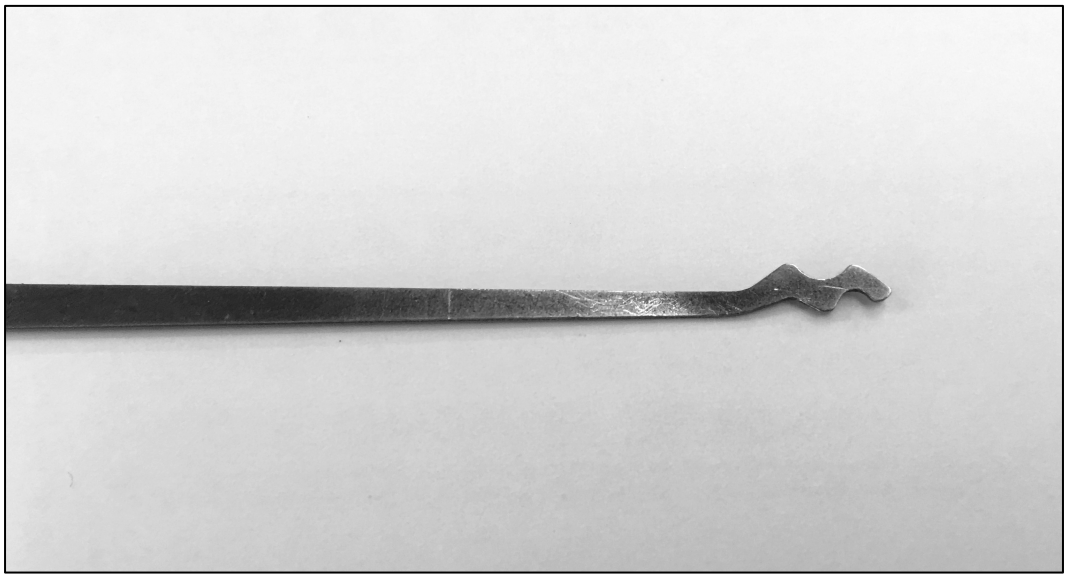


Figure 9.10: A lock-raking pick. A locksmith may open a lock by moving one of these tools quickly over all the pins inside the cylinder, while using another tool to simultaneously try turning the lock. The bumps on the rake will jiggle the pins up and down randomly, in search of the configuration that the proper key would create.

This analogy can take us a little further. Sometimes, when a locksmith rakes a lock, the right combination is not stumbled upon, but in those cases, the locksmith needs only to try again. The causal effects of one pattern upon another are similar, in that we need not expect the first pattern to be guaranteed to have its effect. Rather, we assume it can try over and over again. As long as there is *some* likelihood of its effect, the effect can eventually occur. Eventually, as we develop our discrete analyses, we will give each symbolic structure a *probability* of having its causal effect of producing other patterns in the same environment (like the p mentioned above, in the case of X and Y producing Z; see also Figure 9.8). Those probabilities will be meant to reflect such things as whether or not the causal pattern comes into contact with the patterns it tends to interact with, and whether or not those patterns are oriented in the right direction in order to interact properly when they do come into contact, and so on. And so we can add to this list of probabilistic factors the

ratio—however big or small it is—between the vibrational variations of the pattern that do in fact have the effect we consider the pattern to have and those that don't.[316]

For these reasons, I think it is fair to consider a good many of the states of an atom or a molecule as being organizationally identical to one another (in the sense of containing the same organizational potential—offering dynamics that may reliably interact with other patterns to produce still other patterns). Perhaps not all of a pattern's states are able to have the same organizational effects. But most of them are able to transform into one another, converging to a smaller set, many of whose members have the same organizational effects. I see this as naturally discretizing the world at the molecular level, just as long as we specify that the organizational effects of a pattern are probabilistic.

[316] For now, none of the probabilities we use will need to be realistic. Those kinds of details can be left for later studies. But it is notable that they can be very small. All that will matter for now is the relative scale of these causal probabilities with respect to the probabilities that patterns might be destroyed by random braising energy.

C. Spontaneity

As the wind of time blows into the sails of space, the unfolding of the universe nurtures the evolution of matter under the pressure of information. From divided to condensed and on to organized, living, and thinking matter, the path is toward an increase in complexity through self-organization. . . . Molecular chemistry has created a wide range of ever more sophisticated molecules and materials and has developed a very powerful arsenal of procedures for constructing them from atoms linked by covalent bonds.

—Jean-Marie Lehn (2002)

We have seen that the formation and maintenance of self-organizing systems are compatible with the laws of physical chemistry. We must now confront this idea with the major problem of biology: How did biological systems arise?

—Ilya Prigogine, *et al.* (1972)

The word "spontaneous" means something along the lines of "on its own" or "without external cause". Throughout the remainder of our explorations of identity, I will be referring to the ways that interactions between patterns occur in terms of whether or not those interactions occur spontaneously. I will tend to use examples of chemical reactions; however, the ideas we will be exploring are more general and abstract notions about the construction and destruction of patterns. Chemical systems just make for an excellent domain in which to find examples that are simple, clear, and relevant. Speaking of spontaneity in chemical systems, however, could cause some confusion among scientists, because there is a sense of the term "spontaneous" that chemists have already

appropriated into their technical lexicon in order to describe and categorize chemical reactions, and their use of the word has a very different meaning from mine.

The chemist's usual sense of the term can be thought of as denoting energetic spontaneity. A reaction is *energetically spontaneous* if it does not require an external source of energy for it to occur. A reaction is not energetically spontaneous if it requires some kind of energetic push (over the energy-of-activation "hill") to force the reactants to combine or to decompose.

The sense of the term we will be primarily focused on can instead be called *organizational* spontaneity, and it obviously is more closely linked with the notion of spontaneous organization. A process is *organizationally spontaneous* if it does not require an external source of organization for it to occur. That is to say, if the blueprints for the products of the process lie entirely within the reactants and the environment, then that process, and the production of its products, are organizationally spontaneous. If, however, some part of the organizational potential for creating those products lies in a structure other than the reactants, then the process is not organizationally spontaneous. Those other information-bearing structures are required in order to get the system to produce the products.

The most interesting systems turn out to involve non-organizationally spontaneous (catalyzed) processes, but before we can analyze those more interesting kinds of systems, we first need to understand the features of organizationally spontaneous processes.

Spontaneous Synthesis

Molecules—like other patterns—are constructed by a combination of two general processes, which can be called *synthesis* (the binding-together of two or more patterns into one) and *decomposition* (the splitting of a physical pattern into two or more parts). Studying these processes, and sequences of them, can give us a foundation from which to understand some potential mappings between

patterns and the kinds of organizational graphs that can account for identity. There is a significant difference between the *spontaneous* versions of synthesis and decomposition and the *catalyzed* versions, and we'll find that it is the catalyzed versions that ultimately fuel teleological systems; but let's begin by looking at the simpler, spontaneous varieties since, for starters, the phenomenon of identity still occurs within them and, in any case, they are the underlying processes that give rise to the catalyzed varieties.

Consider first the case of spontaneous synthesis. The process of a proton and an electron coming together to form a hydrogen atom is one example. We can characterize any basic synthesis in terms of two initial organizational structures, A and B, simply binding together to become a new organizational structure, C. That is to say, under some environmental conditions: $A + B \rightarrow C$.

The context within which I would like us to consider these patterns is a logical space in which, at any moment, each of A, B, and C may or may not actually be present. The assumption that any pattern might disappear may seem excessive in cases when our symbols stand for individual atoms, or even subatomic particles, which generally are highly stable. However, the assumption turns out to be harmless in those cases, and yet it becomes necessary when those symbols stand for molecules or larger patterns that might have various susceptibilities to energetic perturbations.

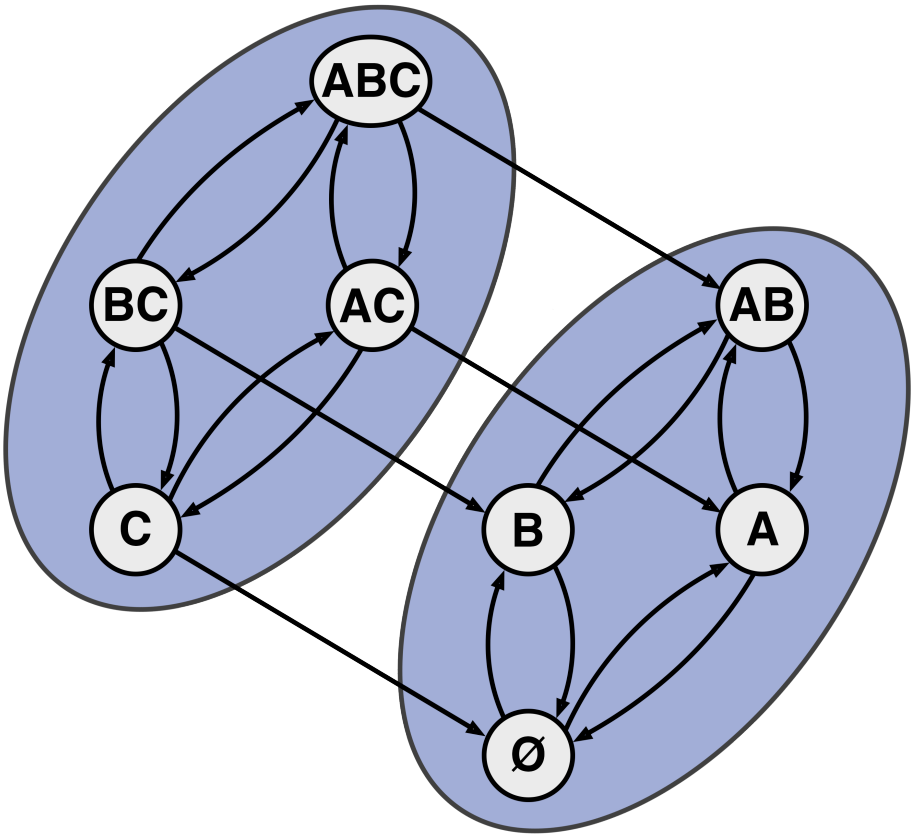


Figure 9.11: A graph representing the potential organization of A, B, and C, in an environment where A and B are known to spontaneously occur (that is, they are freely available via ingress from beyond the bounds of the system, or they are spontaneously produced by other unspecified processes within the system) and where all three of the patterns are subject to potential decay or destruction. Two strongly connected components result from the transitions in this graph. The separation of those SCCs represents the fact that decay in the C dimension will form a one-way transition between two subspaces whose internal structure is specified only in the A and B dimensions. One might think of this graph as being the Cartesian product of three smaller graphs: Ø←→A, Ø←→B, and C→Ø.

The graph in Figure 9.11 represents the potential organizational states of any three patterns, A, B, and C, assuming only that A and B (but not C) are in abundant supply flowing in from beyond the environmental boundary or, perhaps, spontaneously produced from other unnamed processes within the local environment. (Note that we are not yet assuming that C is spontaneously

synthesized from A and B; we will add that assumption to our system shortly.) One can see that each permutation of the set of patterns of interest {A, B, C} is represented as a distinct node in this graph. That is to say that there are nodes representing each of the possible combinations of the three letters: {}, {A}, {B}, {C}, {A, B}, {A, C}, {B, C}, and {A, B, C}. There is also a tangle of transitions between the nodes in the graph, but those transitions fall into two simple categories (up-bound and down-bound) that allow us to easily understand their roles.

First, in order to represent the assumptions that both A and B are freely or spontaneously available in the environment, we allow any state to transition to another state that appends A or B to whatever the original state already had. This can be taken to mean that if actual A or B disappears from this environment, more will soon be produced or introduced (by unspecified but assumed environmental processes). There are eight up-bound transitions that correspond to this notion, and they make up two parallel faces of a cuboid: The Ø state may transition with some likelihood to A, and also to B, and each of A and B may transition with some likelihood to AB. Similarly, C may transition to AC and also to BC, while each of AC and BC may transition to ABC.

Second, the transitions in this graph also represent the fact that the organizational structures of A, B, and C are all potentially subject to braising. The result of the braising of C, for instance, is that any state that contains a C, namely {C, AC, BC, ABC}, has a transition to a state that is nearly identical, but without the C, namely {Ø, A, B, AB}. Likewise, there are four such transitions in the A dimension, and four more in the B dimension. Altogether, these twelve transitions make up a full set of down-bound edges outlining the entire cuboid.

Once the transitions are drawn, analysis shows that this graph consists of two SCCs, each one of which contains four nodes. One of those SCCs represents the existence of C (all of its nodes contain a C), while the other one represents the absence of C (none of its nodes contain a C). The set of one-way, down-bound transitions represents the fact that while some possible states contain

the *actual* organization of C, no state contains the *potential* organization of C. While the potential organization for both A and B exists in the environment (as per our assumptions), C will never be constructed in the system represented here; if some C does unexpectedly come to enter the system, it can only fall apart.

We can augment the model now by adding new transitions that correspond to the spontaneous rule for synthesis: A + B → C. This can be done by drawing an arrow from any node in which A and B simultaneously exist to a new node where both the A and B have "disappeared" and C stands in their stead. This represents the relationship between A and B being knocked across the basin boundary such that they fall into a potential energy well, bind together, and become C. There are only two states where A and B coexist, and so only two arrows need to be added: one from the AB state and one from the ABC state, each terminating on the C state. The modified graph is given in Figure 9.12.

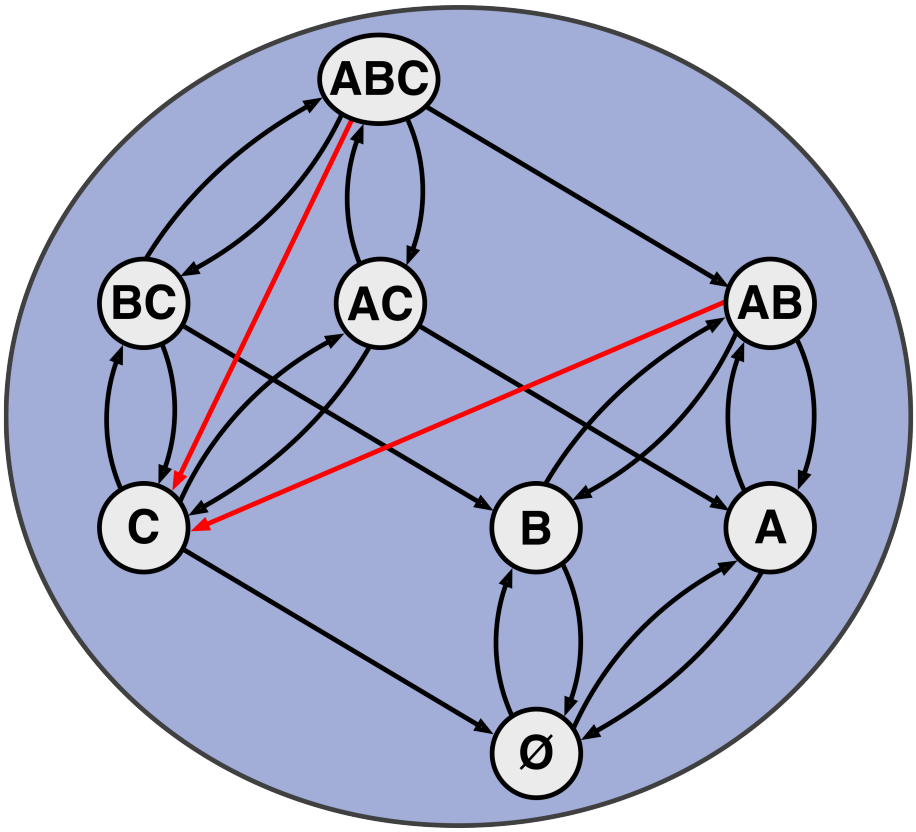


Figure 9.12: A graph that represents the potential existence, within an environment, of patterns A, B, and C, modeling the case of spontaneous synthesis in which A and B are assumed to spontaneously occur, and there is also some chance that A + B → C.

In this new graph, all eight nodes together form a single SCC. In other words, every node in the graph is now organizationally identical. That means that each of the states within the graph contains the same organizational potential (in this environment). Regardless of what actually exists at any particular moment, the potential for A, B, and C all to exist is embedded in the assumptions about the organizational contents of *any* of the nodes. So for instance, when only C exists, our environment has the potential to produce more A and B (that was an assumption of this environment that was true also in the previous graph); when only A and B exist, C can be produced from them, and we can assume the A and B will soon be replaced by more A and B; and even when

none of the three patterns are actually present, there is always the potential for A and B to come about and then to produce C. Each state of the graph has differing *actual* organization; but for every one of those states, the total *potential* organization is equivalent: the potential for all three patterns always exists.

This of course makes sense. Since A and B are spontaneously arising reactants, and since they spontaneously combine to form C, we should also expect C to be spontaneously organizing in the same environment.[317]

A Reduction

If we look back at the two SCCs in Figure 9.11, we can see that the graph represents a situation in which any state with a C can transition—directly or indirectly—to any state without one, but not vice versa. When we add the A + B → C transition, the graph (Figure 9.12) then morphs into one in which any state without a C can also (even if indirectly) transition to any state with one. In both cases—with and without the rule for the spontaneous synthesis of C—the SCCs produced reflect the fact that A and B are really irrelevant here. That is to say, changes in the presence of actual A or B have no effect on the presence of potential A or B. And the reason they are irrelevant is that A and B are present as potential organization in every state of the system. There is no doubt about their organizational presence. They are always there.

In fact, we might even notice that the roles of A and B are much like the roles of their own precursors (let's call them a_1, a_2, b_1, and b_2 for now), which we have already implicitly taken to be environmental assumptions. We have accepted the idea that any time A or B is destroyed, more will

[317] Even if A and B are not subject to braising in our environment—imagine they are fundamental particles, for instance—the graph without those eight down-bound arrows still produces a single SCC wherein all states are organizationally identical.

come about precisely because A and B themselves are spontaneously organizing or introduced in this environment, which entails the fact that their constituents (a_1, a_2, b_1, and b_2) are spontaneously organizing or introduced in this environment. But if we don't need to include a_1, a_2, b_1, and b_2 in our diagram, then, for the same reason, we can also leave out A and B themselves from the diagram. We can logically truncate or reduce both the graphs above in both the A and B dimensions, leaving C as the only element whose existential fate is under consideration.

Another way to think of this is to count A and B as assumptions in the smaller organizational space of just C. Assuming the spontaneous presence of A and B, the organizational rule A + B → C in ABC space represents the same thing as does Ø → C in C space. Working only in C space produces a graph whose general topology is equivalent to the condensation of the graph in ABC space, with one node containing A and B by assumption, but no C, and another node containing A and B by assumption, and also including C. The connection between the two nodes is bidirectional, representing the facts that (i) C is susceptible to braising, and (ii) when A and B coexist, C can be produced.

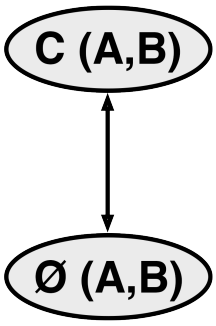
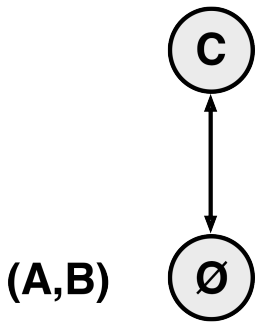


Figure 9.13a: A graph representing the fact that, in this environment, C forms spontaneously. The A and B from which C forms are listed in the parentheses of every node in the graph as a way to make explicit the assumption that those two patterns are always organizationally present, whether or not they are physically present. In other words, this graph represents the same system as Figure 9.12. The two nodes here still form only a single SCC, since C is subject to braising and is also spontaneously reoccurring from the given A and B.



9.13b. A simplified notation for organizational graphs. This is the same graph as in 9.13a. The difference here is that the parenthesized A and B have been removed from the specification of each state, and instead stand in the space around the graph, in order to represent their organizational presence in the environment within which all the possible states of the graph exist. Again, this graph represents the very same system as Figure 9.12. To further simplify our notation, the parenthesized components whose presence is assumed may sometimes be left out.

In general, we can reduce any space we are analyzing with respect to the organizational contents that are available in every state of the space. And we can call those patterns that are ubiquitous in the system "organizational assumptions", and list them as parenthesized organizational contents. Reduction of this sort will be important not only to understanding the organizational roles of the various components in a system, but also to simplifying the spaces we are analyzing.

Spontaneous Decomposition

Decomposition is the opposite of synthesis and so it looks different, but the end result is similar, in that a spontaneous process always results in complete organizational identicality between states containing the precursors and those containing the products of the process. One real-world example of spontaneous decomposition is the reaction that occurs when hydrogen peroxide dissociates into water and gaseous oxygen. The chemical formula for that reaction is $2H_2O_2 \rightarrow$

$2H_2O + O_2$. The more general form of this organizational process can be stated as: P → Q + R (in the context of some environmental conditions).

We can again build up to a graph representing this scenario by first constructing a graph framing its background assumptions within a logical environment in which, at any moment, each of P, Q, and R may or may not actually be present. As before, using such a framework will help us determine which states hold the potential organization for which patterns.

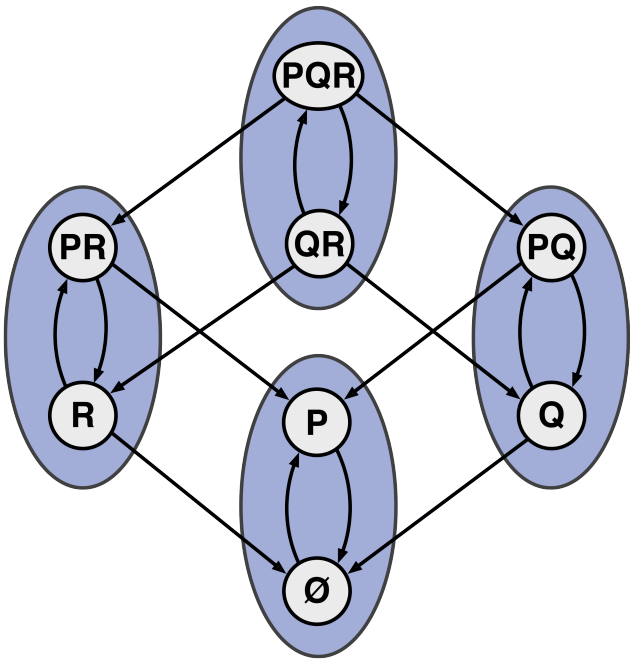


Figure 9.14: A graph representing the potential organization of P, Q, and R, in an environment where P can spontaneously occur, but Q and R cannot. The reader comparing this graph to Figure 9.11 will note that the two graphs similarly contain eight nodes (here labeled in P-Q-R space rather than A-B-C space) but, in order to show more clearly the SCCs that have been formed, the P-Q-R version has been rotated so that the P dimension is upright. Four SCCs are produced by the transitions in this graph. The separation of those SCCs represents the fact that Q and R decay irreversibly. The internal structure of each SCC displays the fact that while P may decay, it may also be spontaneously produced or introduced at any point in time.

The graph in Figure 9.14 represents the potential organizational states of any three patterns, P, Q, and R, analogous to what we saw with patterns A, B, and C in Figure 9.11. In this case, however, we only assume the existence of P (but not Q and R) to be either flowing into the environment or spontaneously produced within the local environment.

We can again walk through the up-bound and down-bound transitions in this graph to understand their roles. The four up-bound transitions—from Ø to P, from Q to PQ, from R to PR, and from QR to PQR—all represent the dependable production or introduction of P within any state where it is not yet actually present. P is potentially present in every state. And just as before, the twelve down-bound transitions represent the fact that any of the organizational structures (P, Q, and R in this case) are potentially subject to braising. There are four SCCs produced by this set of transitions, representing one-way paths of decay from the top SCC (where Q and R both exist) to either of the middle-level ones (where only one of those two patterns exist), and finally to the bottom one (where neither Q nor R exist). In this graph, no state contains the potential organization for either Q or R.

As before, we can now augment our model by adding the transitions that correspond to the spontaneous causal rule P → Q + R. In this case, the arrows to be added travel from each of P, PQ, and PR and all terminate on QR.[318] The final graph for spontaneous decomposition is given in Figure 9.15.

[318] We could also add an arrow from PQR to QR, since the idea here is to follow our decomposition rule and add an arrow from any state with a P to one that removes the P and leaves in its stead a Q and an R; however, that transition already exists, so we can leave it alone for now.

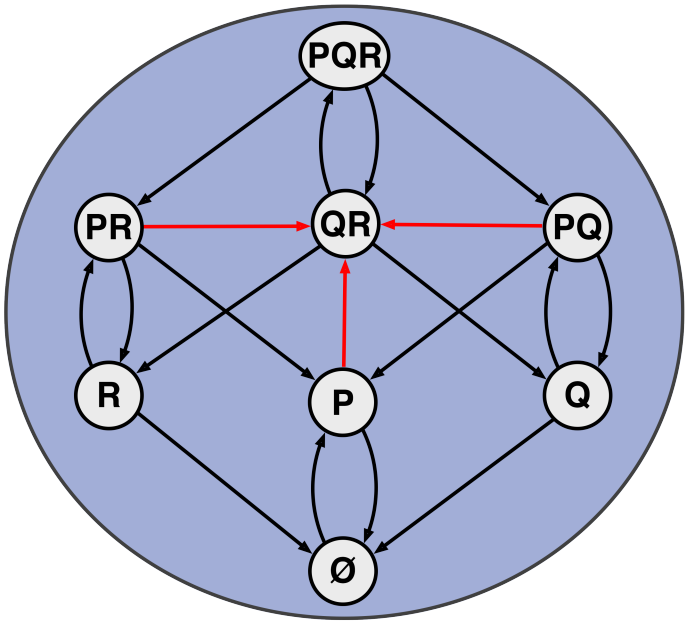


Figure 9.15: A graph that represents the potential existence, within an environment, of parts P, Q, and R, modeling the case of spontaneous decomposition in which P → Q + R. Every node in this graph belongs to the same singular SCC, thus implying that all the nodes are organizationally identical. This makes sense of course: Since we are assuming that P is freely available, and we are assuming P spontaneously forms Q and R, it follows that P, Q, and R are all bound to eventually exist. No matter what the current state of the environment, even if those patterns don't yet exist, the potential for them all does.

Just as we saw in the case of spontaneous synthesis, when we add in the last few transitions that represent the causal rule we are modeling (P → Q + R), all eight nodes of the graph come to form a single, organizationally identical SCC. The organizational potential of every state (in this environment) is the same, with each one containing the potential organization for all of P, Q, and R.[319]

[319] And, as we saw previously with spontaneous synthesis, even if a freely available precursor, such as P, is not subject to braising in our environment—imagine it is a fundamental particle, for instance—the graph without those four down-bound arrows still produces a single SCC wherein all states are organizationally identical.

Other Reaction Types

While our topic is really about the interactions of patterns in general, it can still help to follow chemists in categorizing broad categories of interactions. Chemical reactions are typically grouped into four basic kinds—synthesis, decomposition, single-replacement, and double-replacement—along with a handful of more specialized kinds, including combustion, redox, and some reactions that occur with larger organic molecules (such as rearrangements and pericyclic reactions). We won't take the time to look at all these varieties of interaction in detail because, ultimately, the spontaneous version of each results in the same general picture, but it is worth noticing the fundamental logic surrounding each of the basic forms.

In both single- and double-replacement reactions, two reactants swap parts to produce two products. For instance, in the single-replacement reaction $Cl_2 + 2KBr \rightarrow Br_2 + 2KCl$, the chlorine atoms (Cl) from the first reactant replace the bromine atoms (Br) from the second reactant, leaving two organizationally distinct products. Similarly, in the double-replacement reaction $Na_2S + 2HCl \rightarrow 2NaCl + H_2S$, the sodium atoms (Na) from the first reactant and the hydrogen atoms (H) from the second reactant swap places, again producing two organizationally distinct products. Some such replacement reactions occur by a singular process wherein the two molecules collide, with a kinetic energy above the activation energy for the reaction, and the new atoms slide smoothly into place just as the old atoms slide out. In other cases, reactant molecules may be broken apart (ionized or radicalized) by some prior decomposition, and the new products are then assembled from the parts.

In this case, the organizational processes may be chained together, but in the end they can be organizationally reduced to the same form as the smooth, singular process.[320]

Organizationally speaking, both types of replacement reaction can be characterized by the rule A + B → C + D. And in an environment where A and B are assumed to exist, the organizational graph for reactions that follow this rule results, as with our spontaneous synthesis and decomposition reactions, in a complete SCC that encompasses all nodes. As before, all states of the system contain the same organizational potential and thus are organizationally identical (see Figure 9.16).

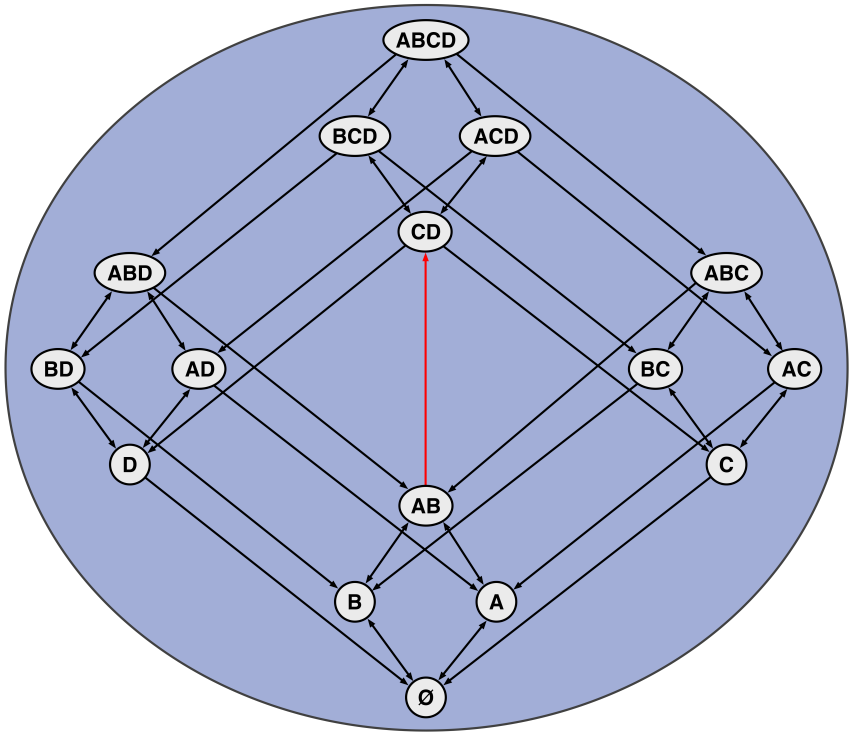


Figure 9.16: The organizational graph that corresponds to single- and double-replacement reactions, following the rule A + B → C + D in an environment where A and B are assumed.

[320] For example, one could combine A → E + F; B → G + H; E + G → C; and F + H → D and wind up with A + B → C + D. The organizational contents of the quasi-stable intermediates (E, F, G, and H) would simply be irrelevant, since whatever is produced would soon be consumed.

Combustion reactions are typically of the form A + B → C (for instance, $C + O_2 \rightarrow CO_2$ describes the combustion of carbon, as in coal) or A + B → C + D (for instance, $2C_2H_6 + 7O_2 \rightarrow 4CO_2 + 6H_2O$ describes the combustion of ethane). In either case, we have already reviewed the organizational results of these formal rules above—the first is synthesis, and the second is organizationally homologous with single- and double-replacement reactions.

Chaining Processes

The spontaneous kinds of organizational processes we have so far been looking at may combine transitively to produce *compound* spontaneous processes, as one might expect. And also as one might expect, when they combine procedurally, their organizational identicality combines as well.

The combustion of ethane mentioned just a few moments ago is an example of this. Although we simplify the reaction to the form A+ B → C + D, in actuality to get seven oxygen molecules to combine with two ethane molecules in the formation of four carbon dioxide and six water molecules, it requires a series of steps with a number of intermediate structures, not always occurring in the same order. If it didn't require this complexity, we would have to imagine the very unlikely coincidence of all seven oxygen molecules colliding with both ethane molecules at just the right angles and velocities all at the very same moment. That of course just doesn't happen.

Another example comes in the form of redox reactions, which chemists often more explicitly separate into individual reduction and oxidation parts, in order to keep track of the electron transfers that occur in the process. For instance, the combustion of magnesium, which is summarized as $2Mg + O_2 + 4e^- \rightarrow 2MgO + 4e^-$ may be separated into its oxidation half, wherein

each magnesium atom donates two electrons, and its reduction half, wherein each oxygen atom gains two electrons:

1) $2Mg \rightarrow 2Mg^{2+} + 4e^-$ (Oxidation)

2) $O_2 + 4e^- \rightarrow 2O^{2-}$ (Reduction)

The third step, after the electrons have been donated from the magnesium to the oxygen, is for the $2Mg^{2+}$ and the $2O^{2-}$ to spontaneously combine into 2MgO. We can encode this reaction with a chain of three spontaneous organizational rules that map onto the above reaction equations: (1) A → B + C; (2) D + C → E; and (3) B + E → F, along with the assumptions that Ø → A and Ø → D (that is to say, the magnesium and oxygen are provided). The result, which is too tangled an organizational diagram to meaningfully show, is a single SCC that encompasses all the nodes in the space of A–B–C–D–E–F. (If we analyze this a bit further, we might notice that the electrons freed in the first step and consumed in the second, and the magnesium and oxygen ions produced in the first two steps and consumed in the third, are all intermediates, and so, for the purposes of organizational analyses, we can ignore these parts and simplify the reaction to $2Mg + O_2 \rightarrow 2MgO$.)

Here is a simpler example of how two synthetic reactions might be chained together: A and B might form some intermediate I, which might then combine with C to form D (A + B → I; I + C → D). If both these processes are spontaneous in this environment, then, in the spontaneously occurring presence of A, B and C, we will find the organizational potential for I and ultimately also for D. This can be shown by producing the full five-dimensional graph of A-B-C-D-I space, with 32 nodes that are all organizationally identical; but it is easier to look at a reduction across the environmentally assumed A, B, and C dimensions, which results in a simple graph representing Ø →

I and I → D. While Figure 9.17 helps us to visualize the intermediate in this case, ultimately that graph also reduces to the merely spontaneous Ø → D.

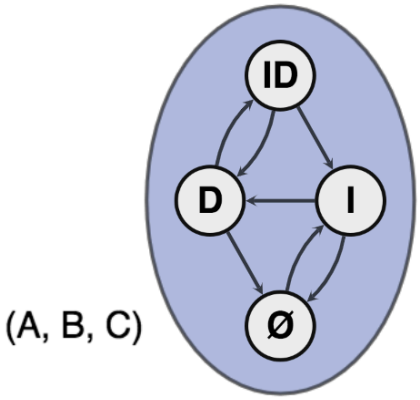


Figure 9.17: An organizational graph representing chained spontaneous synthesis wherein A + B → I and I + C → D, in an environment where A, B, and C are all assumed to exist.

The same can be shown for any combination of chained decomposition and synthesis interactions. As long as the initial reactants are assumed to exist in the environment, and the partial reactions are spontaneous, then the organizational potential of the resultant patterns will tend to be present in every state of the system.

Multipart Processes

We will continue to look at other types of reaction mechanisms just a little further. But the conclusion we are going to reach is that, in some sense, the results of all spontaneous reactions are the same: space-covering SCCs in which every state of the system contains the same potential for all the organizational products.

The spontaneous formation of a product from three or more pieces or the spontaneous decomposition of a pattern into three or more pieces might only ever occur as chained processes of syntheses and decompositions from two parts. Still, to satisfy some curiosity, we can look instead at the idea of direct three-part synthesis and decomposition, and then generalize what we find to multipart synthesis and decomposition (whether or not those activities ever strictly occur). In all such cases, the processes continue to have fully organizationally identical graphs that may reduce across their environmentally available reactants.

For instance, Figure 9.18 shows the four-dimensional graphs for both A + B + C → D and P → Q + R + S. Each of those graphs results in a single SCC encompassing all the possible states of the system.

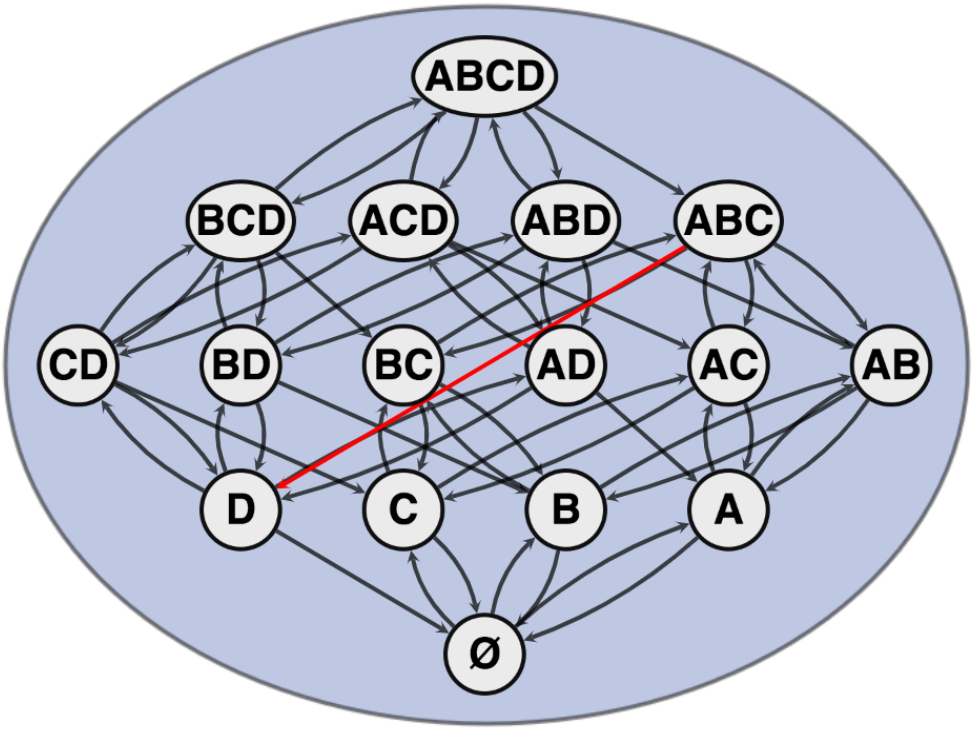


Figure 9.18a: An organizational graph representing synthesis from three parts. The organizationally causal rule represented here is A + B + C → D. I've simplified the graph by including only one of the red-arrow transitions that signifies the causal rule. The remaining transitions in the graph represent the effects of braising and the spontaneous introduction of environmental assumptions A, B, and C. The entire graph forms a single SCC, meaning that all the nodes here are organizationally identical states, the spontaneous organizational potential for D is available in any state of the system, and so the whole graph is reducible to just Ø → D.

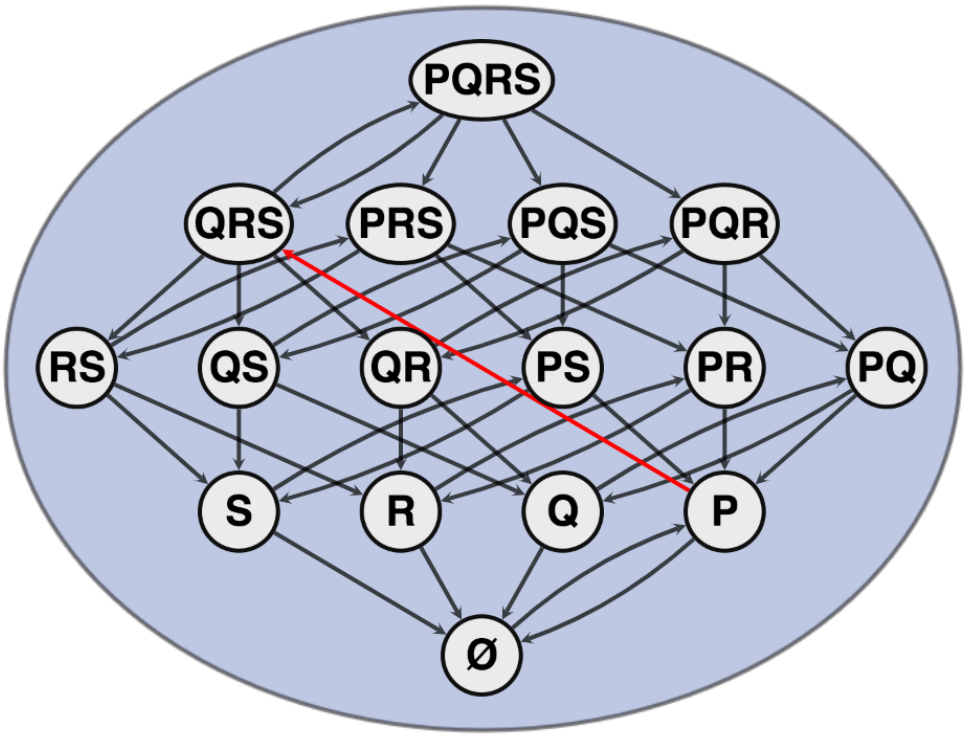


Figure 9.18b: An organizational graph representing decomposition to three parts. The organizational rule represented here is P → Q + R + S. Again, I've simplified the graphs by including only one of the red-arrow transitions that signifies this rule. The rest of the transitions are braising transitions or correspond to the introduction of environmental assumption P. As with the previous graph, all the nodes here form a single SCC together, and so Q, R, and S are spontaneously organizing, and the graph is reducible to Ø → Q + R + S.

In general, any multipart synthesis or decomposition will produce a graph with a single SCC encompassing all states because, in any such spontaneous system, the reactants are assumed to exist, and the products spontaneously arise from those reactants, meaning that the organizational potential for all patterns is present in every state of the system.

The Real World

In the foregoing analyses, I have tried to show how we might account for a broad range of idealized spontaneous interactions between patterns. The real world, however, is a far, far messier place than I have so far admitted. For just one example of the kinds of complications that regularly arise, we might notice that sometimes a pattern may be formed by some process that occurs in one environment, then the pattern is moved to another environment, after which it may serve as a precursor to a secondary process that would not occur in the first environment. The former process might even be prevented from happening in the latter environment (by chemical "poisoning" of the reaction, for instance) so that the entire set of processes not only *does* not but also *could* not happen under identical environmental circumstances. In such cases, a model of the potential states of the system would have to be extended in some way to account for the changing of environmental assumptions within the changing of the states. Certainly, in the case of organisms that are able to move themselves in search of greener pastures, it seems almost overwhelmingly complex to try to apply the simple kinds of analyses we have been looking at.

I feel confident, however, that future work could overcome these kinds of complications, but I am not prepared to work through any significant portion of those analyses at the moment. And I think that trying to make those extensions, while eventually necessary, would at present only distract us from the core issues. The points that I would like to stay focused on now are that there potentially exist, in a system, a set of states that are all organizationally identical because of their mutual accessibility from one another, over time; and furthermore that any model of identity that accounts for sameness over time can provide, first, an account of the existence of the kinds of patterns that might exist and, second, the basis, as we'll see, for an account of the *subject* in subjectivity.

D. Catalysis

It is then shown that several simple and compound bodies, soluble and insoluble, have the property of exercising on other bodies an action very different from chemical affinity. The body effecting the changes does not take part in the reaction and remains unaltered through the reaction. This unknown body acts by means of an internal force, whose nature is unknown to us.

—Jöns Jacob Berzelius (1835)[321]

Spontaneous reactions give rise to spontaneously organizing products. Although we took the time above to spell all that out in lengthy detail, the basic logic is straightforward: Any time a particular environment contains spontaneously arising reactants that also spontaneously combine or decompose, we can expect the products of those reactions to eventually exist. Because of the organizational potential of the assumed reactants, the products are organizationally present even when they are not physically present. We get space-covering SCCs in the organizational graphs of these processes because all the states of a system with spontaneous reactions contain the same organizational potential.

There are, however, forms of synthesis and decomposition that involve additional organizational precursors—catalysts—and that are thus not what we will call organizationally spontaneous, even if they are energetically spontaneous. That is to say, although such reaction sequences may, as a whole, be thermodynamically favorable and will thus proceed forward in an energetically spontaneous manner, that inclination comes about largely thanks to the presence of the organizational potential in the catalyst, the role of which is to give the overall reaction a bit of a

[321] As cited in Lindström and Pettersson (2003).

causal shove in the right direction. In the absence of the catalyst, the foodstuffs that eventually go on to form the products may not contain sufficient organizational potential for those products. In short, catalysts are able to change the total organizational potential of a system, simply by bringing in some of their own.

The role that catalysis plays in affecting organizational potential may not seem too noteworthy at first, but it ends up being rather profound, and the details of why that is the case will become clearer as we continue analyzing the phenomenon. To motivate that analysis, however, it is worth recalling that catalysis—both in its literal, chemical sense and also in its more figurative, causal sense—plays key roles in the majority of the internal behaviors of every living organism we know. In fact, the majority of biochemical pathways, in every known cell type, all follow a series of chains and cycles of catalyzed reactions. And, although the boundary between physical and chemical causation is certainly blurry, many organismal structures—cell membranes, flagella, arteries, bones, feathers, scales, shells, and so on—play less chemically and more physically causal roles that ultimately enable (*i.e.* metaphorically catalyze) various biochemical processes. Organizationally speaking, any physical pattern that produces dynamics involved in the creation of another pattern, but at the same time, does not itself undergo change during that process, can broadly be construed as a catalyst (or, to coin a term, one might say a "causalyst"—a dynamical *causalytic* factor in the formation of a pattern).

The most general fact about catalysts is that they are patterns that, on one hand, play organizational roles in transforming other patterns and yet, on the other hand, are organizationally unchanged by the end of the process. Chemists define catalysts as chemicals that are involved in altering the rates of reactions but that are neither consumed nor ultimately transformed in those reactions. The importance of catalysts remaining organizationally unchanged is that organizational content—information—can be preserved: if a catalyst contains some of the potential organization

for other patterns, but that potential (along with the catalyst's own actual organization) is not spent in the process of producing those patterns, then the overall result is *an increase in organizational redundancy* (for some products) within the system.

For instance, when reactants (say b_1 and b_2) spontaneously combine, they sacrifice their own organization in order to produce that of B. The reactants' potential is transformed—it is spent—in order to create the actual B. In contrast, when A catalyzes the formation of B, the b_1 and b_2 foodstuffs may still be sacrificed, but the catalyst is retained within the system. It is not that more of A may be produced or introduced (as in the case of environmentally assumed patterns); it is just that the very same A is still there in actual form. It is unspent. And since the contribution of that A to the organizational potential of the system also remains, it can go on to catalyze the same reaction again (as long as the organizational potential for b_1 and b_2 are continually produced or otherwise introduced, or if the resultant B is broken back up into its constituent parts by braising).

The ability of a catalyst to increase organizational redundancy may seem merely to be a mildly interesting thing: a small amount can be used, again and again, to produce a large amount of some desired product, which can be very useful in many industrial processes. But catalysis gets to be even more interesting when taken to the next level. As Kauffman has shown us, when arranged in just the right ways, this kind of information-preserving productivity can be exploited so that the patterns that become protected by redundancy are the very catalysts that help to create redundancy. Like self-running multiquines, autocatalytic sets are able to endlessly produce the means to produce themselves and thereby increase their own redundant organization within a system.

There are a variety of reaction mechanisms that correspond to catalysis. Sometimes, a catalyst is merely an adsorbent (sticky) surface upon which other reactants are able to encounter one another. Other times, a catalyst may form a covalent bond—and thus merge, in a chemical sense—with another reactant, only to be released sometime later by breakage of that bond. Still other times, a catalyst may be broken apart entirely and, after each of its pieces has served some distinct causal role in the reaction, the catalyst may then be reconstructed later in the reaction sequence. In more physical (causalytic) cases—such as that of a pair of scissors that helps a paper become divided or that of a magnetic-confinement fusion reactor that helps two lighter atoms fuse together into a heavier one—the causalyst may not bond with the substrate at all, but only provide dynamics—particularly shaped fields of force—that compel precursor patterns to become composed or decomposed into products.

Despite this diversity, we can develop some standard organizational formulae to represent the phenomenon. Ultimately, catalysis is any chained series of spontaneous reactions that happen to result in the eventual production or release of one of the initial reactants (the catalyst) along with some other products or byproducts. We can look at some standard examples first, and then, more interestingly, at an example in which one of the additional byproducts is also the same catalyst, resulting in what we can call either autocatalysis or just chemical replication.

Consider first a classic example of catalysis in which a platinum-group metal is able to catalyze a hydrocarbon reaction. For instance, we can think of the reaction we looked at earlier wherein gaseous ethene (C_2H_4) combines with gaseous hydrogen (H_2) to form gaseous ethane (C_2H_6) in the presence of solid platinum. The primary role of the platinum in this kind of reaction is to stretch the bonds within the reactants by adsorbing them onto its surface, a process that damages

those bonds, and thereby reduces the energy of activation necessary for the reactants to bond to one another. This kind of stretching mechanism is a very common feature in certain kinds of chemical catalysis.

Here is one way we can model the sequence of events that occur in the platinum-catalyzed reaction: We can begin by labeling the ethene A, the diatomic hydrogen B, and the platinum surface C. When the ethene adsorbs to the platinum, it forms a complex (a quasi-stable combination of the atoms in both those reactants) that we can call E. In particular, as the two carbon atoms of the ethene are attracted to two nearby platinum atoms, the double bond between the carbon atoms is stretched to its breaking point, whereupon it becomes a single bond, and the newly freed electrons play a role first in the adsorption that binds the ethene to the platinum, and then later in binding it with the hydrogen atoms to form the new ethane molecule.

If our lump of platinum has a large surface, and if there are a lot of ethene molecules around, then various kinds of Es come and go regularly. That is to say, structurally different Es may come to exist depending on the locations of adsorption onto the platinum, but those Es often have the same causal capacity, with respect to hydrogen molecules, making them equivalently Es for the catalytic process.

Then, when a diatomic hydrogen molecule also happens to adsorb onto the platinum, right beside the site of adsorption of the ethene, we can call the entire structure, which typically lasts only a very brief amount of time but is nonetheless organizationally distinct, F. Lastly, multiple events might ensue from this F, depending how near the new hydrogen atoms are to the adsorbed carbon atoms from the ethene: If they are too distant (call the structure F' instead of F), perhaps the whole complex just sits there doing nothing for a while, until some other moving molecule bumps into the structure, providing some kinetic energy that might release one of the pieces from the surface without their interacting together; in that case, the reversibility of the adsorption typically results in

the original reactants forming once again. If, however, the hydrogen atoms are near enough to the ethene carbons (so we have an actual F), the more likely result is that the hydrogen then bonds to the ethene, and the electrons that were active in the state of adsorption come to be involved instead in the new carbon–hydrogen bonds, thus releasing the new larger hydrocarbon from the platinum catalyst. We can call the resultant ethane molecule product D. By this point, the catalyst, C, has been discharged from its duty and has been released unchanged. The summary of this series of potential events can be given by the following organizational rules:[322]

A + C → E	Ethene + Catalyst → Complex-E
E + B → F	Complex-E + Hydrogen → Complex-F
F → C + D	Complex-F → Catalyst + Ethane

As we can see, our catalytic story is a series of transformational changes that results in one of the initial reactants, C, being reproduced or released in a later stage. In the process, E and F are short-lived intermediate structures while, ultimately, the total change represented here is only that A and B may be combined to produce D. A string of spontaneously occurring events that necessarily involves the catalyst is able to produce some product from some reactants, but also releases the catalyst in its original form. It is of note that this sequence of partial reactions can in fact be reduced to a single organizational rule. But let's wait to look at that formal transformation until after our second example.

[322] We might add another couple of rules to the system, signifying those cases described above where adsorption is followed by desorption and no new product is produced. Those rules would be E + B → F' and F' → E + B. In these cases, the hydrogen B adsorbs to the catalyst complex E, but is not near enough to bond with the ethene in that complex. F' is formed but does not decompose into D and C the way that F does. F' instead may only decompose back into E and B (or even A, B, and C). The catalyst may be released, but no D is produced.

Enzymes are the commonest types of catalyst in the biological world, and so they make for an important illustration. The modern understanding of enzymatic catalysis is what is called the lock-and-key theory, whereby each enzyme has a specific shape that is suited to stretching a particular substrate reactant in such a way that those substrates preferentially react in a way that typically would not have occurred in the absence of the enzyme (Michaelis and Menten 1913). Since our previous example was a case of catalytic synthesis, we can look now at an example of enzymatic decomposition. The standard form for enzymatic decomposition is for the enzyme *E* first to bind to a series of points on the substrate molecule *S* (for instance, a sugar or an alcohol or some other organic molecule) to form an activated complex *A*. This substrate binding is usually made up of some combination of hydrogen bonds and covalent bonds. In the case of decomposition, the combined adhesion of those bonds strains one of the bonds of the substrate molecule until it breaks. We can call the new short-lived intermediate structure that consists of the enzyme combined with the two separated halves of the substrate B. In the end, the bonds holding B together are relatively weak and so a moderate amount of kinetic energy will release the parts of what was once the substrate (we can call these parts C and D), leaving the enzyme E intact and ready to work again. Altogether, the organizational rules for this sequence of events are:

E + S → A	Enzyme + Substrate → Complex-A
A → B	Complex-A → Complex-B
B → C + D + E	Complex-B → Products C and D + Enzyme

Once again, however, the result of catalysis is the same. A series of transformational changes result in one of the initial reactants—this time the enzyme E—being reproduced or released in a later stage while, in the process, a new product (or, in this case, two) is produced from the remaining reactants.

Compressing Over Intermediates

Both of these examples—the platinum-catalyzed synthesis and enzymatic decomposition—involve explicit *intermediates* in the sequence of steps that compose those reactions. Intermediates are structures that constitute a kind of organizational opposite to catalysts. While catalysts serve as a consistent source of potential organization both by introducing organization into the system (as a kind of reactant) and by carrying it onward (by remaining unchanged as a product), intermediates are organizationally trivial—they serve as short-lived carriers of organizational content during transformations, but they neither introduce organization into a system nor carry it on later.

If we look carefully, we find that intermediates are everywhere in organizational processes. Not only are they sometimes explicitly present as products that become reactants in sequences of organizational events, but they also are implicitly present within individual organizational events. In any step of synthesis or decomposition, the change of states from reactants to products occurs not as a discrete leap from start state to end, but as a transition through a continuous series of intermediate states, all of which we tend to ignore as irrelevant because they generally tend to be equivalently bound, and to lead from one to another reliably.[323]

[323] An exception to this occurs at the quantum level, where discrete state changes do occur, as far as we can tell. However, let's stay focused on the pattern interactions that occur at the atomic level and above.

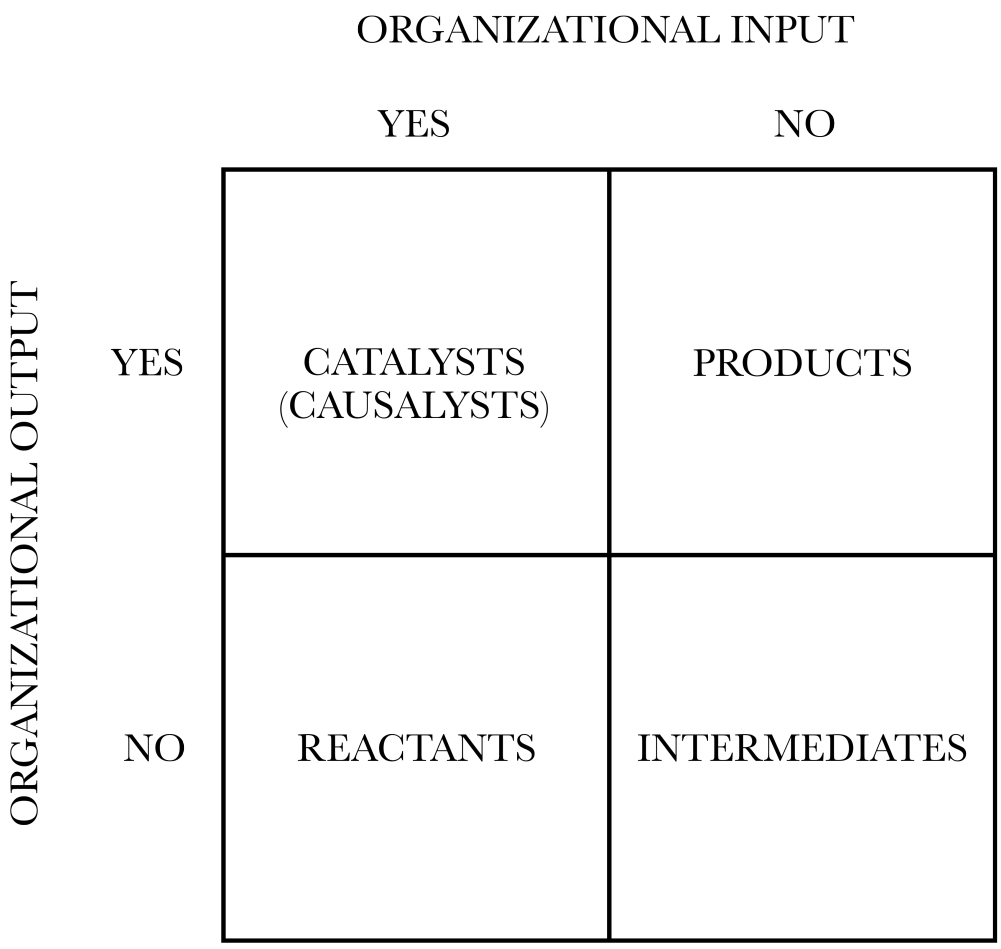


Figure 9.19: A grid highlighting the major classes of structures involved in organizational interactions, organized in terms of whether the organizational potential of those structures is input into the system, output from the system, neither, or both.

Now because of their organizationally irrelevant role, intermediates may be removed from a set of organizational formulae by a process we might call compression. We have already seen a case of potential compression, when we looked at the two chained organizational rules A + B → I and I + C → D. If we postpone reduction of the foodstuffs in this case, we can still simplify the two formula by compressing them together to produce A + B + C → D, leaving the intermediate I out of the picture . . . In short, any time a series of transformations involves an intermediate product

that then becomes a reactant, but that doesn't serve either as an initial reactant or as a final product, the transformations can be combined and, in the process, that intermediate product can be struck from both sides of the new organizational rule. In contrast, any catalyst in the system serves both as an initial reactant *and* as a final product, and so, although it may also appear on both sides of a combined organizational rule, it cannot be struck from the transformation.

Chemists perform the compression of intermediates intuitively in their work, and we can do the same with our two examples of catalysis from above. In the case of our catalyzed synthesis example, our original rules were:

A + C → E

E + B → F

F → C + D

It is easy to see that both E and F serve as intermediates in this set of processes—each of those patterns is a product of one transformation that is later consumed in another transformation. And so one can compress the set of rules across those two intermediate patterns by striking them from each side, and combining the remainders of the rules. The result of that process is a single organizational rule that notably contains neither E nor F, but still contains the catalyst on both sides.

A + B + C → C + D

We can call this organizational formula the *expanded* standard form for catalyzed synthesis. And we will soon use it to develop the *reduced* standard form for catalyzed synthesis and its graphical equivalent.

In the case of our enzymatically-catalyzed decomposition, we had the following series of organizational rules:

E + S → A

A → B

B → C + D + E

If we combine these formulae in the same way as we've just done, striking the intermediates A and B from both sides and then combining the rest, the result is a single rule that shows substrate S decomposing into products C and D, in the presence of enzyme E (which is retained on both sides of the formula).

E + S → C + D + E

Mirroring the names we created above, we can call this organizational formula the expanded standard form for catalyzed decomposition. And we will use it also to develop a reduced standard form for catalyzed decomposition.

Catalyzed Synthesis

It will be easier to graph catalyzed synthesis if we first reduce the standard form the same way we did earlier in our analysis of spontaneous transformations. This just means removing any reactants that we take to be organizational assumptions in the system. So A + B + C → C + D can be reduced across A and B, becoming C → C + D. Graphs representing both formulae are shown in Figure 9.20.

What is important to notice in the case of catalysis is that the system produces an up-bound transition that is not rooted in the Ø node. This reflects the fact that catalysis irreducibly requires one organizational pattern (the catalyst) to produce another (the catalyzed products).

In an organizationally spontaneous system, if there is an arrow from C to CD, it is typically because Ø → D in that environment, and so every state without a D (*e.g.*, C) transitions to a similar state plus the D (*e.g.*, CD). The blueprint for D exists entirely within the reactants that we assume to exist, and that is what makes the product D a spontaneously organizing pattern. But in the catalytic system, part of the blueprint—part of the potential organization—for D lies in the catalyst, and so D is not able to form spontaneously from just any state in the system that lacks it, but only from states that contain C. In other words, the organizational potential for D can be decoupled from the Ø node.

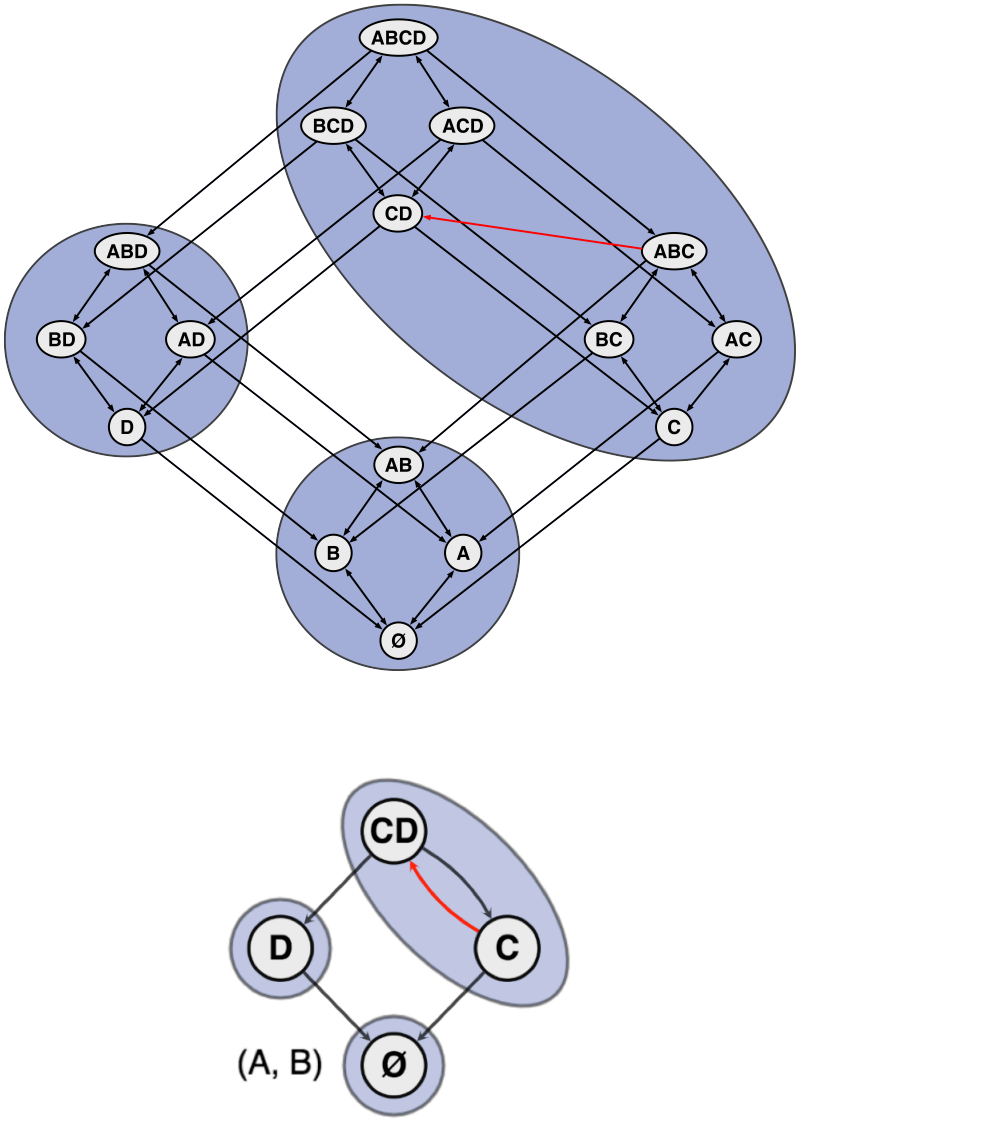


Figure 9.20: Two representations of simple catalyzed synthesis. (a) The full graphical specification of A + B + C → C + D, wherein C is the catalyst for the formation of product D from reactants A and B, which are taken to be environmentally available. (b) A graphical specification of the same process, reduced over the A and B dimensions. We can reduce over A and B because they are taken to be organizational assumptions—widely available foodstuffs—in this environment. As one can see, the condensations of both graphs here are have the same general form.

Decoupling a single pattern from the Ø node does not, in and of itself, allow a pattern to persist longer than it otherwise would. The organizational potential of such patterns is still susceptible to braising, in the same way that a program Q that is written by program P is susceptible to the braising of P (after which no more Q will be made and then it, too, will suffer from random destruction). But this capacity for decoupling that arises from catalysis is the first step toward potentially decoupling the organizational potential for an entire system from the Ø node. If that complete decoupling is possible—and it turns out that it is—then there can potentially exist systems whose organizational identities cannot be constructed from null. Such systems are thus not spontaneous, and yet they may be able to persist for some time by evolving from state to state before collapsing to null—that is, to say, by constructing and repairing themselves.

Catalyzed Decomposition

The case of catalyzed decomposition can also be modeled straightforwardly. The expanded standard form for catalyzed decomposition that we found above, E + S → C + D + E, signifies the way enzyme E helps substrate S decompose into two products without itself being affected. We can use a lettering scheme akin to the one we used for catalyzed synthesis to say the same thing—catalyst C helps reactant A decompose into two products B and D without itself being affected: A + C → C + B + D. And in the usual way, we can reduce this formula over foodstuff A to produce the reduced standard form for catalyzed decomposition, which is just C → C + B + D.

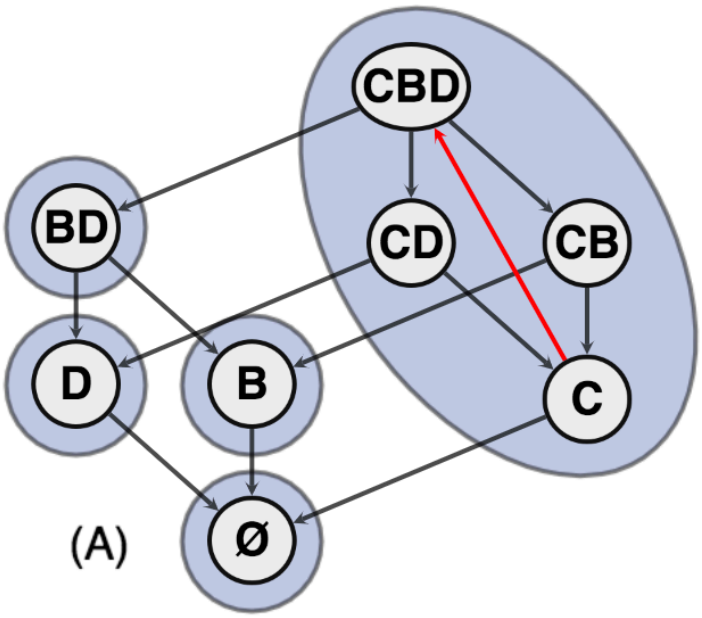


Figure 9.21: An organizational graph depicting catalyzed decomposition. The system represents the rule $C \rightarrow C + B + D$ whereby C catalyzes the decomposition of environmentally assumed foodstuff A into patterns B and D.

We can read the graph in Figure 9.21 as saying that, assuming the provision of foodstuff A, the organizational presence of catalyst C is identical (forms an SCC) with any state that contains both the catalyst and any combination of the two products. That is to say, as long as C exists, B and D are also organizationally present, but the absence of any C implies the absence of the organizational potential for B and D, which then tend only to be susceptible to braising. Again, as with catalyzed synthesis, the production of B and D irreducibly requires the catalyst, and it yields an SCC that rooted in the C node, rather than the Ø node, or in other words an SCC that depends upon C for the blueprints of the other contents, and that is independent of the Ø node for those blueprints.

A Proviso about Time

As I mentioned, chemists normally think of catalysts as molecules that change the *rates* of chemical reactions that are, in any case, bound to occur. In general, this is true, and it is easy to find examples—for instance, the synthesis of rust from iron and oxygen, catalyzed by water and various salts or acids, or the decomposition of hydrogen peroxide, catalyzed by manganese dioxide. Without those catalysts, both processes proceed, but they typically proceed slowly, sometimes taking years to run their course. A mixture of nitrogen and hydrogen gases at room temperature may seem to be entirely stable. Those reactants come to form ammonia so slowly, in the absence of high temperature and a catalyst, that it is as if they don't react at all.

One way to understand this is in terms of the energy of activation required for a reaction to occur. If a certain threshold amount of energy is required for two reactant molecules to break their existing bonds and for their parts to combine into a new product, then, in any given environmental distribution of molecular kinetic energies, once in a (possibly rare) while, a collision that exceeds that threshold will occur, allowing one molecule of product to be created. The question of how slowly or quickly a reaction takes place is the question of how often such energetic collisions occur, which depends on the shape of the distribution of kinetic energies among the molecules. A catalyst provides an alternative source of energy—the potential in the structure of the catalyst itself can do work on the reactants, pulling them apart (partially or wholly) or forcing them together, and thereby making it unnecessary to wait for the kinetic energy of heat to create high-energy collisions. This can substantially speed up a reaction.

For our purposes, however, I don't think a mere *quantitative* change in rate is the most productive way to think of catalysis, even though that is exactly what is going on. For one thing, when a catalyst changes the rate of a reaction significantly, even if the products of the reaction are

the same as in the non-catalyzed version, the result can still be *qualitatively* likened to a discrete change of states. It is as if something that was not able to happen is suddenly now able to happen. For instance, if an energetically spontaneous reaction forms some product, but environmental braising destroys that product (or that product migrates out of the system) at a rate that far exceeds the production rate (or if the production rate is just glacially slow), then, in this system, it is as if the product were not being produced at all. The product does not accumulate or come to exist in any manner that can have ongoing causal effects of its own in the system. If, however, a catalyst is then introduced and the same reaction takes off at a much greater clip, now exceeding the braising or emigration rate, then the product not only will be produced, but also will accumulate, and will be able to have its own effects felt elsewhere in the system. The result is that the presence or absence of the catalyst effectively produces—or approximates—a more black-and-white situation in terms of the existence of the products.

For another thing, larger, more complicated molecules may not react in a particular manner at all, without the presence of a catalyst. The activity of enzymes and the substrates they operate upon in biological systems are the most prevalent examples of this. For instance, think of the decomposition of a lengthy molecule. Imagine our molecule is made of six parts, bonded in a fairly linear fashion: A-B-C-D-E-F. And imagine the weakest link in this molecule—the bond most likely to break upon a random energetic impact—is the one between B and C. In general, whether this molecule decomposes quickly or slowly, its decomposition results in A-B and C-D-E-F. Alternatively, if there is an enzyme that binds to the substrate strongly at C and at E, causing some strain between them (and if, say, the C-D bond is weaker than the D-E bond), then the catalyst will tend to help this molecule decompose instead into A-B-C and D-E-F. That now easy and common decomposition reaction would not have happened at all in the absence of the catalyst, and so D-E-F, for instance, would not have been a product of the system.

The same thing occurs quite obviously in the physical, causalytic interactions between macro-level patterns. Many of the effects these patterns have on one another are vastly unlikely to occur except in the presence of the precise causalytic pattern, or some equivalent (but also rare in the scheme of things) pattern. Random events simply cannot cause coins to be minted, knives to be sharpened, or paper cranes to be folded. It seems as if, with an increase in scale—perhaps of size or perhaps of complexity—there is a shift away from events being determined to occur by the statistics of heat energy, and toward their being determined more discretely and conditionally by the *very specific* dynamic, energetic patterns that either are or are not able to have those effects.

In light of this, as we go forward, I will at first continue analyzing catalytic processes as if they were simply *necessary causes* of the reactions they catalyze, ignoring for a while the view of them as merely rate-changers in processes that would proceed nonetheless. That will be sufficient for understanding the cases of larger more complex patterns—probably from the scale of enzymes and DNA on up to the entire macroscopic world. Later, however, after we've developed a rough working understanding of autocatalytic systems (and when time becomes more important to us, as we begin to analyze the notion of value), we can adjust our models and reintroduce a low probability of spontaneous formation in parallel with a high probability of catalyzed formation. We'll discover that, when we do so, the qualitative behavior of systems does not differ significantly.

Autocatalysis

In certain cases, the product of a catalyzed reaction may turn out to be the catalyst itself. The organizational rule that corresponds to this *autocatalysis* is much like that for any other catalysis: When catalyzed by C, some foodstuffs (say, A and B) might combine to form . . . more C.

$A + B + C \rightarrow C + C$

The basic idea is fairly straightforward. One of the Cs on the product side of the rule corresponds to the catalyst (which is why it also showed up on the reactant side of the rule); the other C on the product side is the new product, which just happens to be the same pattern as the catalyst. When reduced over foodstuffs A and B, we obtain what we might call the standard form for autocatalysis.

$C \rightarrow C + C$

This formula looks curious in that there is only one symbol involved in the entire thing, but that is no mistake. Autocatalysis occurs when a thing helps to make more of itself, and so, really, another way of thinking about the phenomenon is in terms of replication—an autocatalyst is a replicator. We'll look more at the notion of replication in a moment, but first we should take the time to draw an organizational graph of this rule.

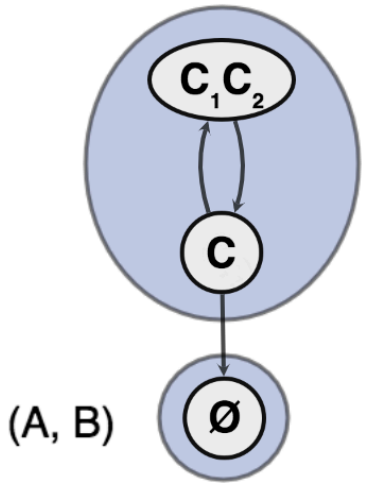


Figure 9.22: An organizational graph corresponding to the autocatalytic formula C → C + C. Here, C_1 and C_2 both represent *the same* C, but it is easier to indicate the replication (the formation of two distinct C) by using multiple symbols. The individual nodes for C_1 and C_2 are unified as just a C node, from which there is an up-bound arrow to the C_1C_2 node indicating that any C can produce another. The graph is also characterized by a full set of down-bound arrows that correspond to potential braising effects. Here, C_1 and C_2 form the simplest version of an *autocatalytic set*, a notion that we will analyze in more detail shortly. The result of that set is a phenomenon we'll see more of soon: all the upper nodes form a single SCC, together; the only node excluded from that SCC is the Ø node. In this environment, any C carries the blueprints for the endless production of more C, while Ø (the absence of C) carries none of the organizational potential for the patterns whose production we are interested in.

This autocatalysis marks an important waypoint as we move forward with our analysis. It gives us the first instance of an SCC, the blueprints of which are contained within the SCC itself, thereby divorcing it from dependence upon the Ø node. It is thus the first semblance of a thing both creating and being created by itself.

Extended Autocatalysis

Earlier we looked at reducing and compressing systems to single rules to make their analysis simpler. Now, I'd like to do the opposite, in order to make it clear that a very wide—possibly infinite—class of extended reaction mechanisms may all be equivalent to the autocatalysis we just looked at. This is important, because these kinds of reaction mechanisms are widely characteristic of biological systems.

We need just one abstract example to understand the potential diversity. We can take the following set of rules and think about their reduction across foodstuffs and compression across intermediates. The result will be the same C → C + C that we already looked at, but the processes involved in the full set of rules afford a glimpse into more complex processes such as those of membrane growth and division, template replication, maintenance of the core components of metabolic systems, and even binary fission of an entire cell. In short, a given molecule or polymer may grow, via a series of steps, to eventually consist in a doubled version of itself, at which point it may also divide into two. The entire cell, one might notice, seems to be a coordinated and interleaved set of such processes.

C + A → I	Autocatalyst + Foodstuff-A → Intermediate-I
I + B → J	Intermediate-I + Foodstuff-B → Intermediate-J
J + D → K	Intermediate-J + Foodstuff-D → Intermediate-K
K → C + C	Intermediate-K → Two Identical Autocatalysts

We can think of C, here, as the autocatalyst, since it plays a role in initiating the cycle and also is doubly present at the end. As the set of rules progresses, that initial C grows and becomes

intermediates I, J, and K, through the sequential addition of foodstuffs A, B, and D. When, at last, all the ingredients have been combined, it turns out that K is a molecule that really consists of two Cs that are joined in some easily dissociable way. K may then decompose by splitting down the middle and turning into those two Cs, resulting in the autocatalysis (*i.e.*, replication) of C.

This basic kind of autocatalytic growth and division underlies all biological growth and division. In the next main section, we'll see at least three distinct variations on this theme: each of the subsystems of Gánti's chemoton is composed of a series of reactions that results in the autocatalysis of some ingredient.

Autoproductive Sets?

Before settling into our upcoming analysis of autocatalysis and its variations, we might try to imagine a parallel space of potentially persistent patterns that we could call "autoproductive sets". The idea would be to find a set of patterns (say, A, B, and C again) that spontaneously change into one another in some sort of cycle, without involving catalysis. For instance, molecule C might join with a_1 to form A, and A might join with b_1 to form B, and B might join with c_1 to form C.

$C + a_1 \rightarrow A$

$A + b_1 \rightarrow B$

$B + c_1 \rightarrow C$

This kind of autoproductive set is, however, strictly impossible, because each formula in the set describes an additive process, and so there is no way that C, for example, could turn out to be the

small kind of thing that could serve as a part of A (since, after all, A has already served as a part of C).

We might instead try to amend this by imagining a scenario in which one of the reactions is a decomposition; perhaps B loses some part, to become C again, with a byproduct of c_1.

$C + a_1 \rightarrow A$

$A + b_1 \rightarrow B$

$B \rightarrow C + c_1$ (where c_1 and C differ from b_1 and A)

But in this case, we again find ourselves stymied. The C that goes into the system in the first step comes out eventually, in the third step, and thus really turns out to be a catalyst. Patterns a_1 and b_1 are consumed as foodstuffs (they are not produced here), c_1 is the sole product (it is produced but not introduced), and both A and B serve as intermediate structures that are re-consumed within the set of processes. Again, what we are imagining here turns out not to be the autoproductive set we'd intended to develop, but just a case of basic catalysis as we've already analyzed it.[324]

Oscillating and Reversible Reactions

Next, we might try to draw some inspiration from empirical discoveries. Chemists have discovered what they call oscillating reactions, which are also sometimes called "chemical clocks" because they appear to change states periodically (see, *e.g.*, Belousov 1959; Briggs and Rauscher 1973; Zhabotinsky 1964). Some of these reactions are visually quite striking, as they display different

[324] One might go a step further, and try to suggest a new product D (instead of c_1) is produced from B, in the hopes that D could then convert to A, B, or C. But if one works out the details, it turns out that this plan results only in a system in which that final A, B, or C turns out to be an autocatalyst.

colors in their different phases. And they often have been cited as examples of spontaneously organizing (but not necessarily teleological) systems in which a kind of order emerges from a simple chemical system (Deacon 2013; Juarrero 1999; Prigogine and Stengers 1984).

Figure 9.23: The Belousov-Zhabotinsky reaction—a kind of "chemical clock"—proceeding in a petri dish. As time goes by, the rings of color emanate, as the blue turns red, and the red turns back to blue. This occurs because the first stage of the reaction sequence produces (differently colored) products that are then consumed in the next stage, and vice versa. In stirred versions of these chemical clocks, the entire solution may alternate abruptly from one color to another. Photo taken by Ted Kinsman.

As it turns out, however, in such reactions, the oscillation that occurs is not an alternation between the existence of some reactant and that of some product, but instead an alternation between relative concentrations of intermediates. The reaction is still ultimately directional in its search for chemical equilibrium, and so it eventually comes to an end, resulting in the final assembly

of a set of new products from the original reactants, rather than some kind of potentially infinite alternation between a series of patterns.

Another example of how the foodstuffs of a set of reactions may transform into one another in a kind of circular fashion, without involving catalysis, is in reversible reactions (Berthollet 1803). While every reaction is in principle reversible, many reactions turn out to be irreversible in practice.[325] Still, there are some reactions that are reversible in practice, and these reactions tend to flow simultaneously in both directions, homing in on an equilibrium point where the rates of the two reactions are balanced. In such a system, two reactants may come together to form a product (or two) that might later decompose (or react) again, and thereby reproduce the original reactants.

Viewed this way, reversible reactions appear to be autoproductive sets. But let's look a bit more closely. We can choose two ways to graph such a system. In the first, we can include (that is, not reduce over) the assumed reactants. That is to say, we can include Ø → A and Ø → B, for instance, along with:

A + B → C + D

C + D → A + B

In that case we get the graph in Figure 9.24. The result is a full space-covering SCC. It is nothing more than a spontaneous system in which A, B, C, and D are expected to come to exist. We have some guaranteed A and B, and the causal rules ensure that this system also guarantees the existence of some C and some D. This is just what we imagined an autoproductive set to be, in the sense that the reactants have the capacity to produce one another without catalysis. But it is not what we had

[325] This happens because the energy of activation for the reactions in each direction may intersect at different points with the distribution of kinetic energies in the environment, leaving the two reaction-directions with probabilities of occurrence that can be several orders of magnitude apart.

hoped an autoproductive set to be, in the sense that we saw above with autocatalysis, in which a set of patterns support one another's existence independently of the Ø node.

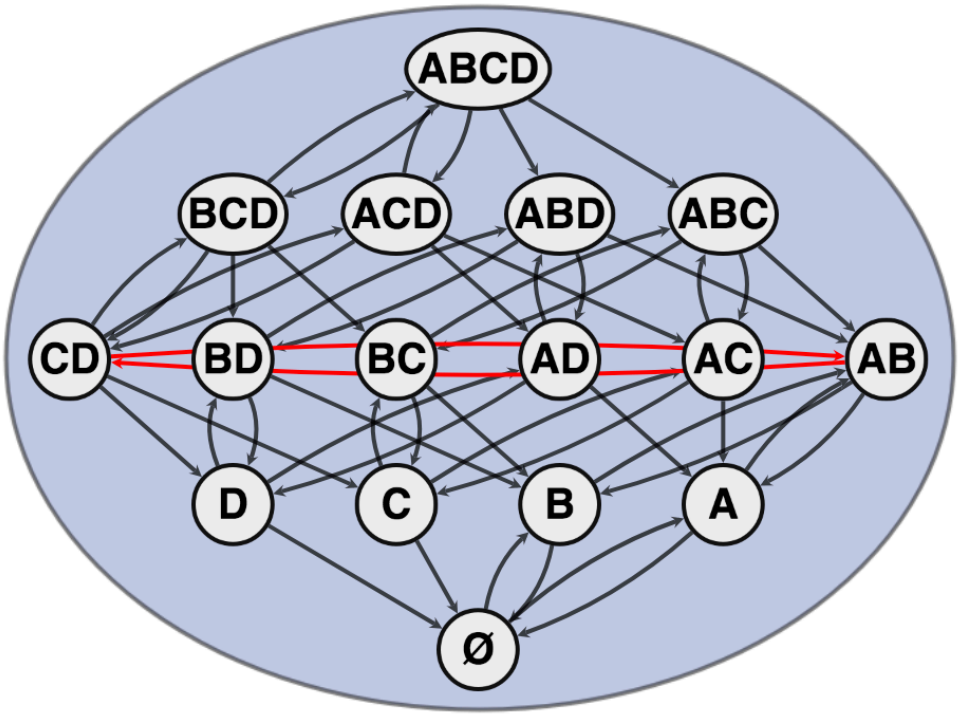


Figure 9.24: An organizational graph of a system in which A and B are assumed (any state without one or the other of these may transition up to a state with the missing pattern) and in which AB and CD are interconvertible states—each potentially producing the other (albeit with possibly differing transition probabilities, leading to the possibility that each of the states is occupied for differing periods of time). There is just one SCC in this graph, encompassing all the states and indicating that the organizational potential of the system is always identical.

The second way of graphing this system might lend a bit more insight into the system. Let's treat A and B the way we did with the autocatalyst C, above, to see what the fate of these patterns is if they are not guaranteed to exist, but if some may just happen to potentially exist. In that case, we get Figure 9.25.

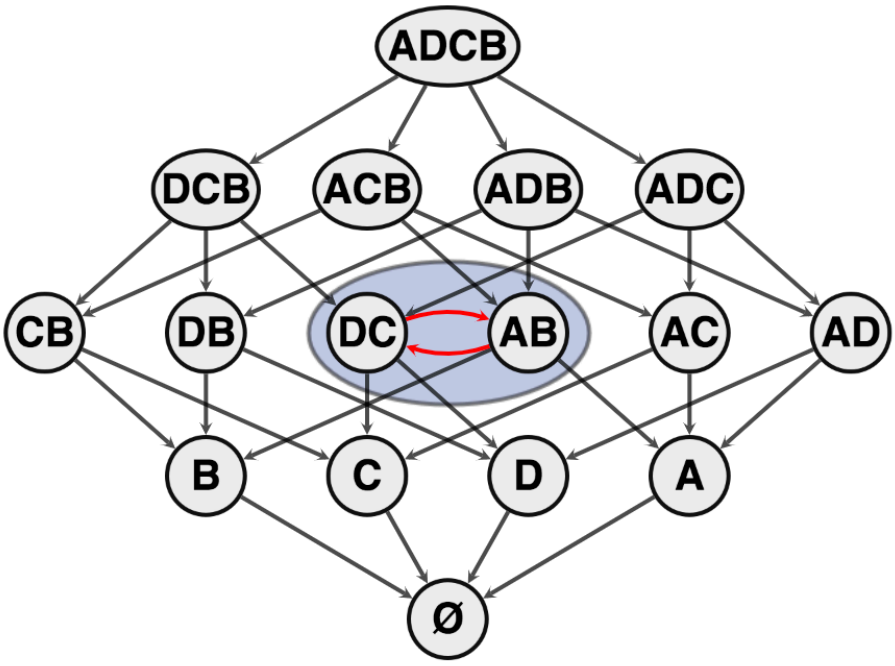


Figure 9.25: An organizational graph representing a reversible reaction system in which A + B and C + D are able to result in one another. The node names here look different from those in Figure 9.24 only because the nodes in the graph have been rearranged so that the AB and CD nodes are next to one another. Some transitions have been left out, for clarity, but they do not affect the SCCs that are produced. There is no guarantee here that any of the reactants come to exist. Nonetheless, if we assume that either some A and B or C and D exist from the start, then our system begins in one of the two states in the marked SCC in the middle of the graph, and it might, for a while, bounce back and forth between those states along the red arrows. What is of note, however, is that this alternation is not bound to continue for any longer than if the reversible reactions had not existed in the first place—once any of the patterns involved decays along any organizational lines, the state of the system moves irrecoverably toward Ø.

In this graph, the nodes AB and CD are organizationally identical with one another, but neither one is identical with any other state. As one can see, the system does not produce a graph that in any way resembles autocatalysis. There are no upbound transitions that might help to

produce either a wider or deeper SCC; the system is equally vulnerable to the braising effects that threaten each individual pattern.

In the end, it seems that autoproductive sets that act like autocatalytic systems are neither theoretically predicted nor found to exist empirically. As far as I am aware, there is no such thing as an autoproductive set. In producing some kind of organizational persistence beyond that of spontaneous systems, it is catalysis and autocatalysis that really matter.

E. Little Miracles of Self-Reference

In the end, we [. . .] self-inventing [. . .] mirages are little miracles of self-reference.[326]

—Douglas Hofstadter (2007, p. 363)

We can try to put the ideas and tools we developed above into action in a more interesting way now by using the graphical system I've been developing to represent the various kinds of autocausal structures for which I set out to develop it.

At this point, we will work only on representing the *identities* involved in autocausal systems, with our sights set on noticing the qualitative differences between the shapes of the SCCs formed in these systems and those formed in our previous analyses. In our upcoming theory of *value*, we can begin to more interestingly quantify those differences as we begin to more deeply analyze time.

We'll begin here with what is perhaps the most straightforward kind of autocausal set, and one that follows immediately from where we just left off with our analysis of catalysis: Kauffman's autocatalytic chemical sets.

Autocatalytic Sets

We can use the same example of autocatalysis that I described on pages xx-yy, in which chemical species A catalyzes the production of chemical species B, and B catalyzes C, which in turn

[326] I've abused this quote from Hofstadter, to some degree. What *he* was talking about was personal, psychological, perceptual identity—a sense of self—while what *I* am talking about is biological, teleological, existential identity. However, the analogy between these two concepts is so compelling to me that I couldn't resist borrowing the relevant parts of his phrase. Both are claims about a kind of identity, and both are based in a self-referential, self-constructing pattern. Hofstadter's complete, original sentence goes like this: "In the end, we self-perceiving, self-inventing, locked-in mirages are little miracles of self-reference."

catalyzes A. To keep the example organized, we'll assume a system or environment wherein six precursor chemicals (a_1, a_2, b_1, b_2, c_1, and c_2) are freely available, and we'll assume that a_1 and a_2 combine by some straightforward type of catalyzed synthesis to form A, and likewise for the combination of b_1 and b_2 to form B, as well as that of c_1 and c_2 to form C. In Kauffman's own notation, this arrangement of chemical and catalytic roles can be drawn as in Figure 9.26.

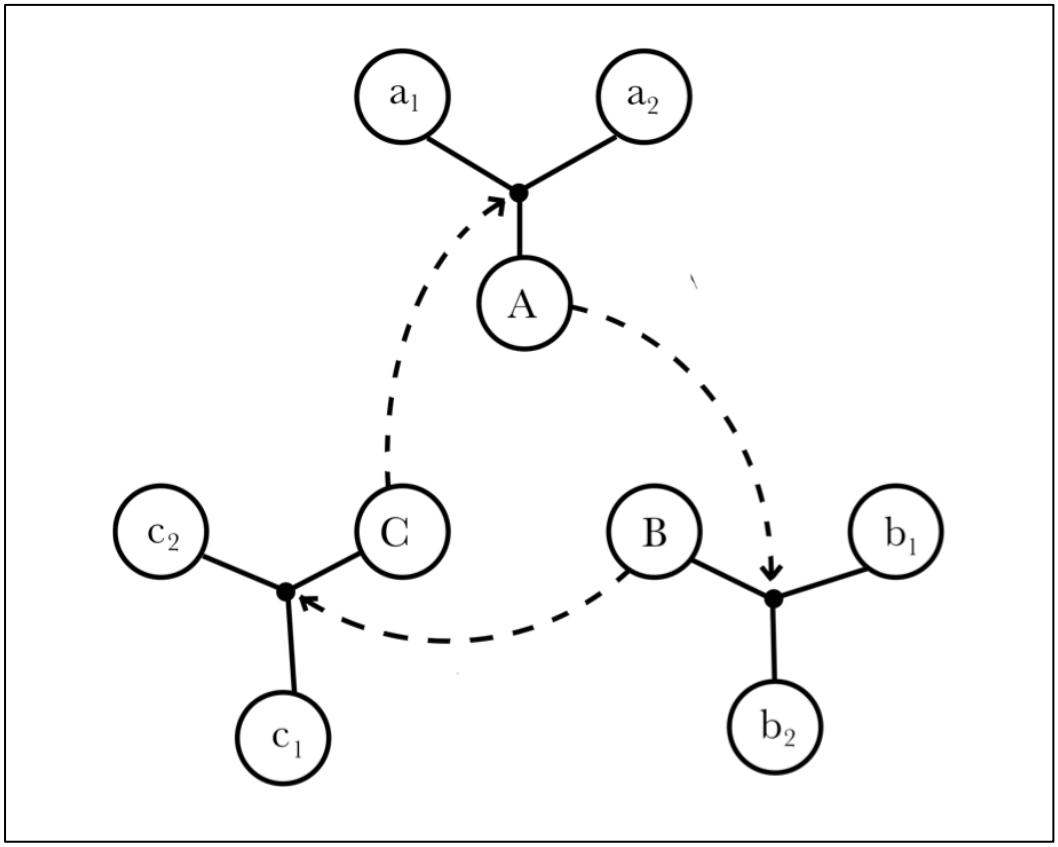


Figure 9.26: An autocatalytic set wherein A catalyzes the synthesis of B from b_1 and b_2, B catalyzes the synthesis of C from c_1 and c_2, and C catalyzes the synthesis of A from a_1 and a_2. The set is drawn using Kauffman's notation, as seen also in Figure 8.2. In that notation, reactions are small black circles, and dotted arrows represent catalytic influences on those reactions. (Note: the notation is ambiguous in terms of distinguishing reactants and products, but one should be able to sort out the details.)

Here are three rules for catalytic synthesis that, together, correspond to the autocatalytic system we are discussing.

$a_1 + a_2 + \mathbf{C} \rightarrow \mathbf{C + A}$

$b_1 + b_2 + \mathbf{A} \rightarrow \mathbf{A + B}$

$c_1 + c_2 + \mathbf{B} \rightarrow \mathbf{B + C}$

As before, we can sideline the reactants that are assumed to exist in every state of the system here, in order to focus more on the causalytic roles of the patterns produced within the system. If we do this, then the organizational rules we are really interested in are just the bold-faced portions of the list of rules above.

The interesting thing about the graph that represents these rules (Figure 9.27) is that its condensation (the graph formed by its SCCs instead of its nodes) differs from the condensation of the purely spontaneous system wherein A, B, and C form, uncatalyzed, from their constituent parts. In the spontaneous case, there is only one SCC, encompassing all the nodes. Here, there are two: the top one, consisting of seven nodes, in which A, B, and C are all organizationally present, and the bottom one, consisting only of the Ø node, in which none of A, B, and C are organizationally present.

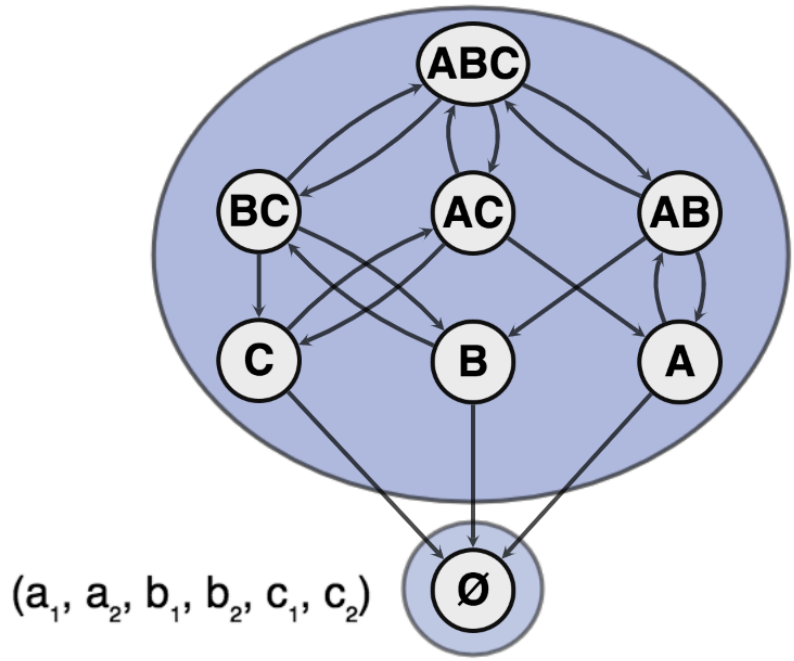


Figure 9.27: A graph representing the autocatalytic set wherein A catalyzes the formation of B, B catalyzes the formation of C, and C catalyzes the formation of A. The top seven nodes are organizationally identical to one another, each containing the organizational content for all nine chemical species {a1, a2, b1, b2, c1, c2, A, B, and C}, in the sum of its actual and potential organization. The bottom node is organizationally distinct from the rest, as it contains the organizational potential for all six of the assumed species, but neither the actual nor the potential organization of A, B, or C.

Although this case has been simplified from reality (because, for the moment, catalysis is being taken as all-or-nothing causality, rather than a change of reaction rate), what emerges from the comparison of this graph with its spontaneous analogue is a qualitative difference between systems. In the spontaneous system, A, B, and C are guaranteed to exist; in the autocatalytic set, as Kauffman's definition makes explicit, if the three components exist, it is only because of their mutual support for one another.

Hypercycles

We can use our organizational graphs to describe a minimal version of Eigen's hypercycles, in a manner similar to what we've just done with Kauffman's autocatalytic sets. If you recall, in the abstract, a hypercycle is just a cyclical autocatalytic superset of individual autocatalysts. And since an autocatalyst is just an autocatalytic set of one, a hypercycle is, really, an autocatalytic set of autocatalytic sets. It seems obvious, then, that the coupling within a hypercycle will only result in another higher-level autocatalytic set. If all the members of each individual autocatalytic set support one another, and if the sets somehow support one another also, then it seems as if the totality of the two (or more) coupled sets should support the existence of all the members of all the individual sets. This is mostly true, but there is some nuance to how it works, depending on the mechanisms by which the couplings are made. We can explore two possibilities for now; however, there are further possible types of coupling, exploration of which will have to be left for later work.

Imagine that A spontaneously converts to a (say, by the sequential addition of some environmentally available atoms or molecules), which then spontaneously converts to two A (by cleavage). And imagine that, in much the same way, B also spontaneously converts to b, which also tends to spontaneously cleave and become two B. In this system so far, both A and B are proliferative autocatalysts by way of some short-lived intermediates.

Now, for our first example, imagine that these two autocatalytic structures are coupled by way of A being a *necessary* co-catalyst for a process in the production of B, and vice versa. In this case, we have actually imagined that A and B are no longer independent autocatalysts, because each depends on the other to catalyze its production. Together, they do form a kind of autocatalytic set, but neither one is self-sufficient any longer. The organizational rules that correspond to this system

would look like the following (wherein each A may stand for each of A_1 and A_2 and each B may stand for each of B_1 and B_2):

$A + B \rightarrow A + A + B$

$B + A \rightarrow B + B + A$

As we noted in the caption to Figure 9.22, autocatalysts are also replicators, which may serve to cause the endless production of themselves. If A helps to produce AA, then (assuming a healthy and wealthy environment) there is nothing stopping each A from helping to produce even more, and thus eventually to yield AAA, AAAA, and so on. Consequently, if A and B are an autocatalytic set, with each pattern catalyzing the production of the other, then when we begin in state AB, where both patterns exist, the potential clearly exists also to move to states such as AAB or ABB . . . and, as with the individual autocatalysts, this proliferation of the two patterns can go on and on and on, leading to a graph with infinitely many nodes containing vast permutations of A and B, all of which are organizationally identical.[327]

In order to graph the two rules corresponding to this kind of coupling, we need to potentially have two As and two Bs in each node. One way to graph that is in the 4-space of {A_1, A_2, B_1, B_2} (where A_1 and A_2 are really the same thing, as are B_1 and B_2), and to graph it with the identical nodes unified. Doing things this way produces the graph in Figure 9.28.

[327] One early reader of this manuscript has suggested that the image that this discussion brings to mind resembles the autocatalytic set in Figure 8.2. Let's not let the fact that both are constructed of series of As and Bs confuse us here. In Kauffman's diagram the series of *a*s and *b*s represent long polymers made up of monomers a and b. Each circle in his diagram represents a single chemical species. In the current discussion, each node that contains multiple As and Bs is meant to represent the simultaneous, side by side existence of patterns A and B, in duplication as many times as specified (i.e. AAB means there are two As and one B).

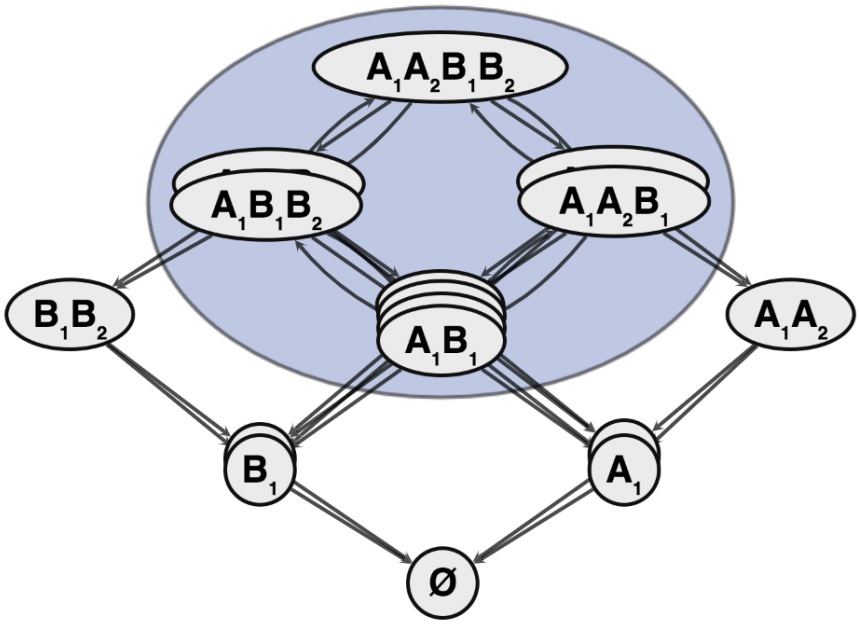


Figure 9.28: An organizational graph of a hypercycle formed by couplings that represent necessary co-catalysis. In this system, B catalyzes some process in the formation of two As from a single A and, likewise, A catalyzes some process in the formation of two Bs from a single B. In the graph, nodes whose edges overlap represent identical nodes; they are meant to be unified, since A_1 and A_2 are really the same pattern, as are B_1 and B_2. What emerges from this form of hypercycle is a single SCC at the top, containing four nodes {ABAB, AAB, ABB, AB}, each of which contains the organizational potential for ABAB, and thus is a potentially proliferative state. The states at the bottom, outside that SCC, all contain either just As or just Bs or Ø. Another way to think about a case like this is that the patterns involved are able to replicate in one another's presence, but unable to do so in one another's absence.

In a hypercycle with this kind of coupling, as long as there is both some A and some B, the system will have the non-spontaneous potential to maintain itself for a while, without either one being spontaneously produced in the system. Although the SCC does not encompass all the nodes above the Ø node, it nonetheless forms a kind of autocatalytic set of its own—one that does encompass all the nodes above the left and right trailing edges of the graph (*i.e.* Ø *and* {A, AA, AAA,

AAAA . . .} and {B, BB, BBB, BBBB . . .}). It is a case in which A and B *together* form a replicator that can resist braising.

Alternatively, we can imagine coupling our two autocatalysts in a way that leaves them independently capable of catalyzing themselves. One way to do that is to have each pattern serve as an *alternative* catalyst (instead of a *necessary* co-catalyst) for the other. In this case, the organizational rules that correspond to our system would look like the following:

A → A + A	A is an autocatalyst.
B → B + B	B is an autocatalyst.
A → A + B	A also catalyzes the production of B.
B → B + A	B also catalyzes the production of A.

The first two rules here correspond to each of A and B each being autocatalytic; the latter two rules correspond to their independently being catalytic of one another. When graphed, the four rules together produce the system in Figure 9.29.

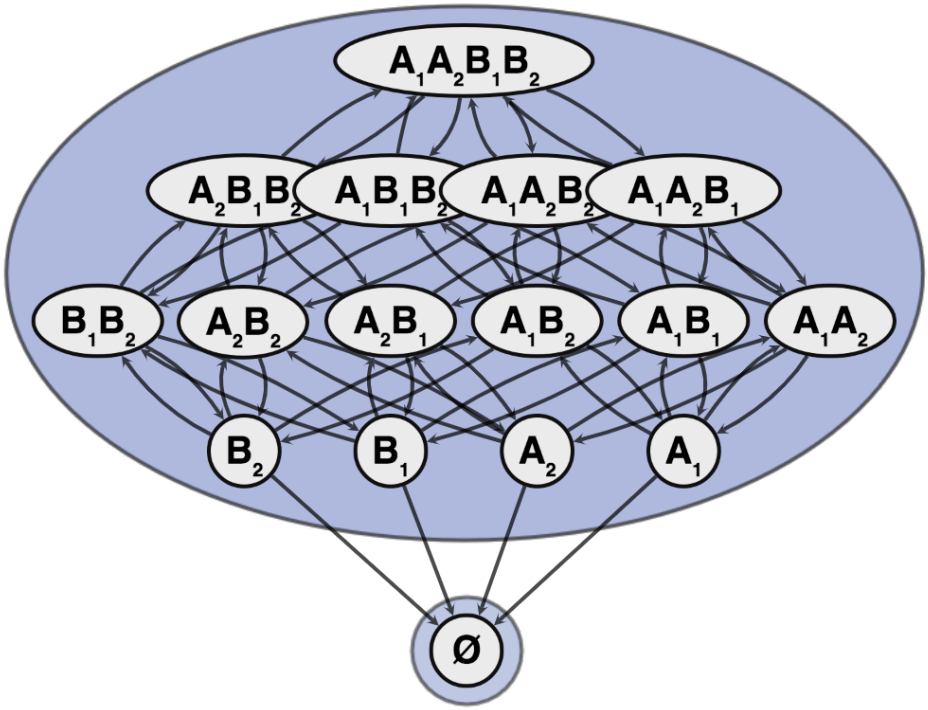


Figure 9.29: An organizational graph of a hypercycle formed by couplings representing disjoint reflexive catalysis of A and B. One could unify identical nodes in this graph, as we did above, but either way, the SCCs formed will be the same. The graph resembles standard autocatalysis, wherein all the nodes except the Ø node are organizationally identical with one another.

What we find in the graph of a hypercycle with this latter type of coupling is that a single SCC encompasses all the nodes of the space except the Ø node, meaning that any amount of either A or B is enough to give the system a shot at producing plenty of A and B, but in the absence of both A and B, the system is unable to produce any.

In summary, any kind of hypercycle is really a complex autocatalytic set. Depending on the couplings used, we get slightly different forms of autocatalysis, but nevertheless each of these systems consists of a set of patterns that are able to *maintain* one another's existence despite the system's being unable to spontaneously *generate* the patterns.

While I am summarizing, I'd like to emphasize one more point: We have learned that autocatalysis just is replication—each process is the production of two copies of a thing from one, which is distinct from multiple copies of a thing being spontaneously produced from just the environment. Spontaneous generation and autocatalysis are both ways of potentially producing progressively more of a pattern or set of patterns, but they do so by qualitatively different methods.

The Chemoton

At this point we may also re-characterize the schema in Gánti's chemoton in terms of organizational potential. Looking at this example in detail will show that a complex model of vitality—albeit one still far simpler than real biological cells—can be clearly outlined as an identity, just as autocatalytic sets and hypercycles can.

Like the systems we have already looked at, the chemoton is a set of organizational contents that, by way of mutual causal dynamics, are able to maintain one another's existence in spite of the braising effects of the environment. As one explores the chemoton model, one may notice that it is arbitrarily extensible, and thus it really represents a broad class of potential models that may be used to approximate various types of cell-biological activities (real or imagined) and their organizational identicality.

We can follow Gánti's schematic for a minimal chemoton, roughly as it was shown earlier. As we noted then, the chemoton has three main subsystems that are coupled to one another. Each subsystem is itself meant to be an autocatalytic system, but none of them is entirely self-sustaining, as each requires causalytic assistance of various sorts from the others.

Our analysis of the chemoton won't be given in terms of the organizational graphs that we have been using so far. Although in principle we could sketch out such a graph, doing so becomes

quite unwieldy in practice because, as the number of components in a model grows linearly, the number of nodes in its graph grows exponentially. But luckily, since we now understand the general principles by which such graphs operate, we no longer need to explicitly draw the graphs out in full in order to understand the shapes and behaviors of the relevant SCCs within them.

We can look first at the metabolic subsystem of the chemoton, which secures materials (and energy) from the environment, and then releases wastes back into it. We can mirror Gánti's metabolic system by using the five organizational formulae found after the diagram below. Together, these rules track a process loosely analogous to metabolic pathways such as the citric-acid cycle in real biological cells. (Please note: this version of the model extends Gánti's by assuming that the conversion between various A patterns is mediated by enzymes rather than spontaneous. We will discuss both cases.)

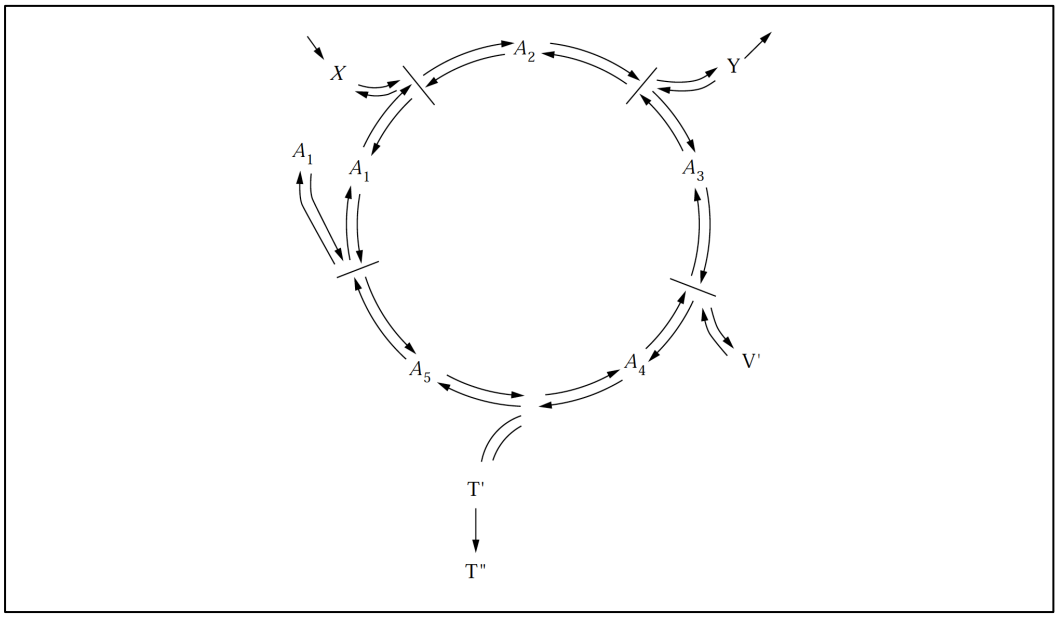


Figure 9.30: The metabolic subsystem of Gánti's chemoton, excerpted from the remainder of his diagram. This system requires the availability of X, discharges Y as waste, and T' and V' as byproducts.[328] It also has a series of molecules that we might call a "backbone" for its semi-stable central role in the process. That backbone goes through transitions from A_1 through A_5, at which point it will have grown large enough (and become properly organized) such that it may split into two identical A_1s, amounting to one form of autocatalysis. Figure adapted from Gánti (1997).

$$T_m + E_1 + A_1 + X \rightarrow A_2 + E_1 + T_m$$

$$T_m + E_2 + A_2 \rightarrow Y + A_3 + E_2 + T_m$$

$$T_m + E_3 + A_3 \rightarrow A_4 + V' + E_3 + T_m$$

$$T_m + E_4 + A_4 \rightarrow A_5 + T' + E_4 + T_m$$

$$T_m + E_5 + A_5 \rightarrow A_1 + A_1 + E_5 + T_m$$

[328] Note: the terms "waste" and "byproducts" here are arbitrary assignments when the subsystem is viewed independently; but the distinction between them makes sense once one has coupled the subsystem with other systems that happen to use the "byproducts" but not the "waste".

In the first rule in this set of transformations, enzyme E_1 and the membrane T_m together catalyze the formation of A_2 from A_1 and food sources $X = \{x_1, \ldots x_n\}$. Then, also catalyzed by a membrane structure along with enzyme E_2, that A_2 is converted to A_3, releasing wastes $Y = \{y_1, \ldots y_n\}$. Gánti seems to be using some shorthand in his modeling when he combines all of the food sources into one symbol as if they are provided in just one step, and all the wastes as if they were produced and released in another single step. Whether or not this is realistic with respect to biological cells is not important to us because, as we saw earlier in our organizational graphs, foods and the elimination of wastes can be taken to be environmental assumptions, and thus they can be sidelined in our analysis; our graphs will always reduce across those dimensions.

Next, A_3 converts to A_4 and, in the process, releases a byproduct called V', which (just as we saw with X and Y) could as well stand for a series of products $\{v'_1, \ldots v'_n\}$, each similarly produced at different stages within the same or a larger cycle or, even, by a series of organizationally analogous cycles.

We can also imagine expanding Gánti's model by adding one or more steps very similar to this one in which, as one version of A converts to another, some additional product, N, is manufactured. What I am imagining here should be thought of as analogous to the steps from the citric-acid cycle whereby D-isocitrate is dehydrogenated to α-ketoglutarate, which is then further dehydrogenated to Succinyl-CoA, producing two NADH molecules (from NAD+). In the real biological case, those NADH—analogous to the N that I am imagining—will then typically go on to help phosphorylate ADP to form the coenzyme ATP, which is the readily consumable transporter of energy that acts as a common fuel for so many other biochemical processes.

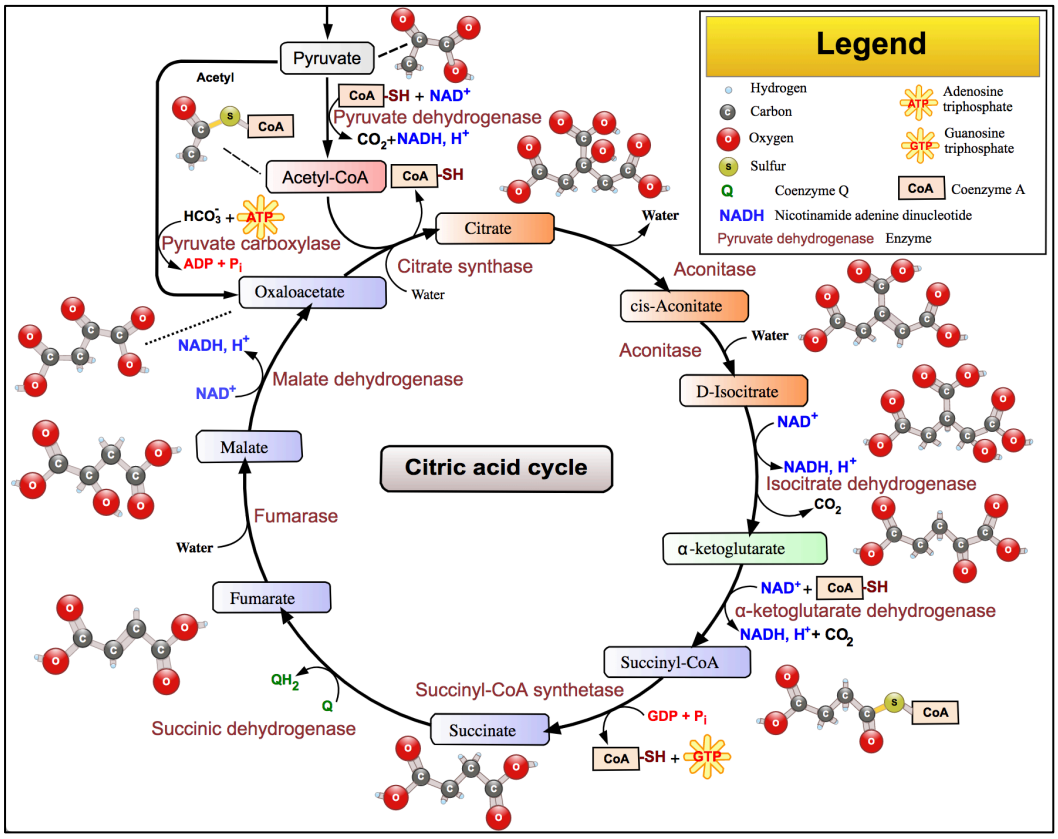


Figure 9.31: A schematic of the citric-acid cycle, showing how a backbone molecule that begins at the top as citrate goes through a series of changes, ultimately returning to its citrate form through the final addition of an acetyl group that can replace atoms previously shed from the backbone during the rest of the cycle. In the process, wastes (*e.g.*, CO_2) and byproducts (*e.g.*, CoA-SH) are produced, along with products GTP, NADH, and $FADH_2$. (Image courtesy of Wikipedia user: Narayanese. Reprinted under the Creative Commons ShareAlike 3.0 Unported license. Downloaded from: https://commons.wikimedia.org/wiki/File:Citric_acid_cycle_with_aconitate_2.svg.)

Either way, the production of N could be analogous in its organizational process to the production of V', and so we can follow Gánti in leaving it out, but we can understand its biological role as being organizationally mirrored by that which we are already analyzing with V'. In short, *something* more may be produced by our metabolic subsystem and consumed elsewhere in the supersystem (the entire chemoton), and this is a sustainable process, thanks to the ingress of some

energetic molecules or photons from the food source X, and to the autocatalytic cycle in which A_1 progresses through a number of stages and ultimately becomes two A_1s, burning free energy from the food sources along the way to produce the organization of that cycle's products.

In the following step of the metabolic subsystem, A_4 is converted to A_5 producing a byproduct, T', in much the same way that V' (and N) were produced. And by the time we reach the last step, Gánti seems to imagine, the sequential addition of various members of X has caused A to grow into a molecule—A_5—that may now be cleaved in two, resulting in an identical pair of A_1 molecules.[329] Through this process, we not only have a factory that produces T', V', and N but we also have a consistent source for the autocatalytic backbone molecules of the metabolic subsystem, as long as X is in supply and the rate of decay of $\{A_1, \ldots A_5\}$ is slower than the turnover rate by which A_1 is duplicated.

Overall, in order to operate, this schematic metabolism requires:

(i) the catalytic capacity of the membrane structure;

(ii) an environment that is able to freely provide the food sources X (which represent both material and energy), and that is also able to absorb or discharge any accumulation of the waste products Y; and

(iii) (potentially) the local provision of a set of enzymes $E = \{E_1, \ldots E_n\}$, which includes those enzymes we were explicit about above, and any that might play other roles.

In Gánti's original form, requirement (iii) is omitted and the system forms an almost independently autocatalytic set. The only catalytic factor not accounted for there is the membrane. In fact, if one

[329] This differs from the citric-acid cycle, which only returns a single citrate by the time the cycle completes. However, our primary focus here is the chemoton. We are only using the citric-acid cycle as an analogy to show that it is energetically possible in the real world to have a cyclical arrangement of anabolic and catabolic steps that produce energetically and organizationally useful byproducts.

imagines the cyclic backbone materials, A_1, . . . A_5, to have come to exist within an artificial membrane near a source of the right spontaneous sources of foodstuffs [say, in mineral micropores near organically-productive alkaline hydrothermal vents under the sea (see Lane 2009; Sojo *et al.* 2016; as well as Goldschmidt 1952; Miller 1953)] then we have an autocatalytic set already.

Alternatively, if the processes of our chemoton's metabolic subsystem are enzyme-mediated, as I've been imagining, then in order to fulfill the definition of an autocatalytic set, the metabolism will need to be coupled with other subsystems that produce both the membrane and that set of enzymes. And aside from trading those gifts for this subsystem's products (V' and T', and possibly N), those other subsystems should be otherwise organizationally self-sufficient within the same environment. Let's continue to look at Gánti's model to see how that commerce plays out.

Gánti intended the information subsystem of the chemoton to be analogous to, but far simpler than, the system in biological cells by which DNA is able to duplicate itself as well as to catalyze the production of other molecules used elsewhere in the cell. We can again try to more-or-less follow Gánti's notation here.

The organizationally causal rules we can derive from his diagram are the following:

$$T_m + pV_n + V' \rightarrow pV_nV_1 + T_m$$

$$T_m + pV_nV_{\ldots} + V' \rightarrow pV_nV_{\ldots} + T_m$$

$$T_m + pV_nV_{n-1} + V' \rightarrow pV_nV_n + T_m$$

$$T_m + pV_nV_n \rightarrow pV_n + pV_n + T_m$$

$$T_m + pVn + E_6 + X + N \rightarrow pVn + E_6 + R + T_m$$

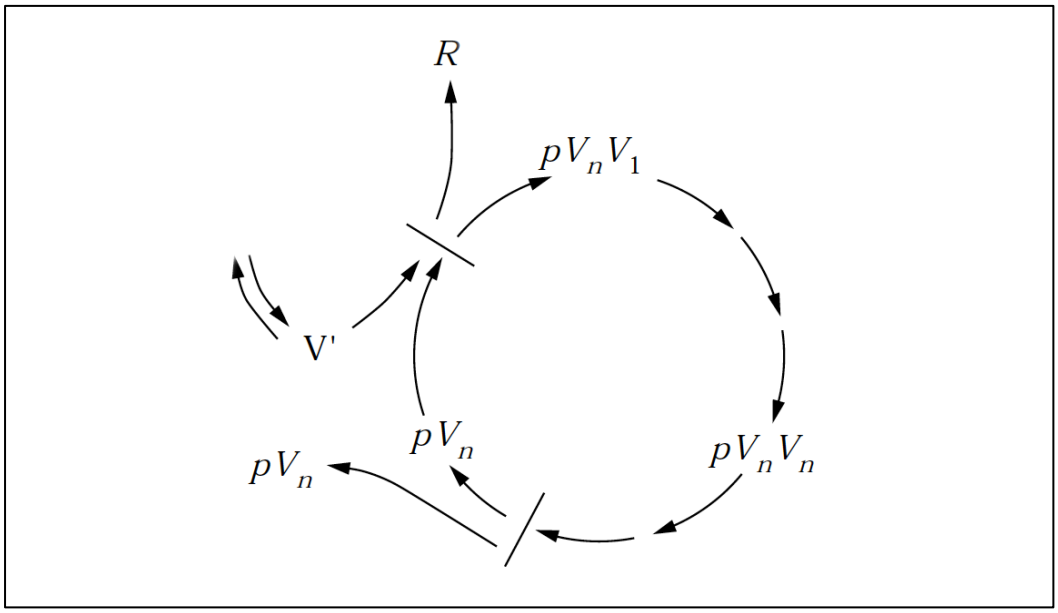


Figure 9.32: An excerpt of the information subsystem of the chemoton taken from Gánti's (1997) diagram. This system requires the availability of V', and produces R. It may also be tasked with the production of E = $\{E_1, \ldots E_n\}$ (please see the main text). The main operation of this information subsystem differs only slightly from that of the metabolic subsystem. In the metabolic case, there was growth and division of a backbone molecule, but the method of growth was not specific; in the information subsystem we have a case of *template reproduction*, whereby a polymer made of *n* sub-molecules grows a parallel copy of itself through a coordinated process of pairing of the sub-molecules (much as RNA and DNA do in their replication by sequential nucleotide addition). V', here, stands for some small set of usable submolecules $\{v_1, \ldots v_m\}$, which may be combined in some linear fashion to form polymer pV_n. As the cycle proceeds, pV_n grows to pV_nV_1 and so on, until it becomes pV_nV_n, after which the doubled polymer unzips to form two distinct pV_n molecules. Figure adapted from Gánti (1997).

The way Gánti intends the symbols in this cycle to be understood is that any symbol starting with a "p" stands for a polymer, made of somewhere between 1 and 2*n* monomers V, in much the way that RNA or DNA is made of a string of nucleotides. As with RNA and DNA, Gánti's polymer templates are able to pair up according to an organizational schema that allows the content

of the template strand to be duplicated in the copy (however, unlike RNA and DNA, which use conjugate pairs, Gánti's system assumes like-pairings that produce a direct copy, rather than a complementary copy). This process of piecewise pairing is represented by the first, second, and third organizational rules above (although the second rule—the one with polymers subscripted with ellipses—stands in for an unspecified number of analogous steps, depending on the number of monomers in the polymer).

To make this process clear: if Gánti's pV_n stands for a template molecule such as LMNOP, then some of the various pV_nV_i would look like the following.

pV_nV_1	pV_nV_2	. . .	pV_nV_{n-1}	pV_nV_n
L	LM		LMNO	LMNOP
|	| |		| | | |	| | | | |
LMNOP	LMNOP	. . .	LMNOP	LMNOP

By the end of the third organizational rule, the template has become fully duplicated (pV_nV_n) and, when the transformation denoted by the second-to-last rule occurs, the doubled template simply unzips its two halves to produce two identical pV_n (each of which is an individual LMNOP).

In addition to the process of template replication, Gánti suggests that each pVn in the information subsystem plays another role, in the production of a molecule R, which itself serves as a precursor to, or enzyme in, the membrane-production subsystem. (Note: even though Gánti's diagram (Figure 9.32) seems to combine this process with the first step of the polymeric replication, I've separated the production of R from the replicative process, as one can see in the last rule in the list above.) The process of R production in the chemoton can be thought of as analogous to the complex processes of transcription and translation in biological cells (together known as the central dogma of molecular biology). To keep our analysis more simplified, however, we can imagine that

this R production corresponds to some stretch of polymer pV_n simply serving as a catalytic surface (acting as an enzymatic binding site) that helps two widely available foodstuffs (from the set X) to be synthesized into R (if we need to, we might even imagine that general process being helped along by another enzyme—call it E_6—which we could take to be analogous to a ribosome).[330]

We can also imagine that various other stretches of pV_n serve in the same way, along with the same E_6 perhaps, as catalysts for the production of the various enzymes, E, whose organizational sources we have so far left unspecified.[331] The additional rule for this, given below, mirrors that for the production of R above. If the polymer is long enough, many and varied enzymes, each of whose organizational potential is carried by the replicating polymer, can be produced by such a process.

$$T_m + pVn + E_6 + X + N \rightarrow pVn + E_6 + E + T_m$$

The commerce that couples the information subsystem to the metabolic subsystem consists in the provision of intermediate foodstuffs V' and N in turn for E. Both subsystems still require a membrane to indirectly catalyze all of the reactions that occur in the cycle by keeping the various parts of the system near enough to one another to work, and so on, but the two subsystems are otherwise cooperatively self-sufficient, in terms of the maintenance and provision of their core components (A and pVn; and perhaps E and N, if we assume these additions to Gánti's model).

Lastly now, we can turn our attention to the subsystem for membrane formation, maintenance, and replication. Gánti has envisioned this system also to be a nearly self-sufficient autocatalytic process—in fact, the membrane can continue existing (and even replicate) just as long

[330] Real ribosomes are made of a combination of RNA strands and proteins—both encoded for by DNA, but we can leave those details aside for our chemoton model.

[331] What is more, because of the extensible nature of the polymer that bears organizational information (with each segment potentially serving as a different catalyst), this subsystem can, at no further cost, produce any number of further necessary catalysts (even those required for its own more complex construction and maintenance, if we were to draw those requirements out in more detail).

as it is provided with inputs R, from the information subsystem, and T', from the metabolic subsystem. As we'll explore more below, if R and T' come instead from the environment then such membranes will form spontaneously and instead will more or less be what scientists call micelles (see pp. 92–94). Following Gánti, the model for this subsystem consists of two processes that are easier to analyze individually. The first of the two processes is the formation of T monomers from R and T', and the second is the spontaneous organization, and potential division, of a set of causally equivalent membrane molecules, $\{T_m, \ldots T_{m+m}\}$ from those T monomers.

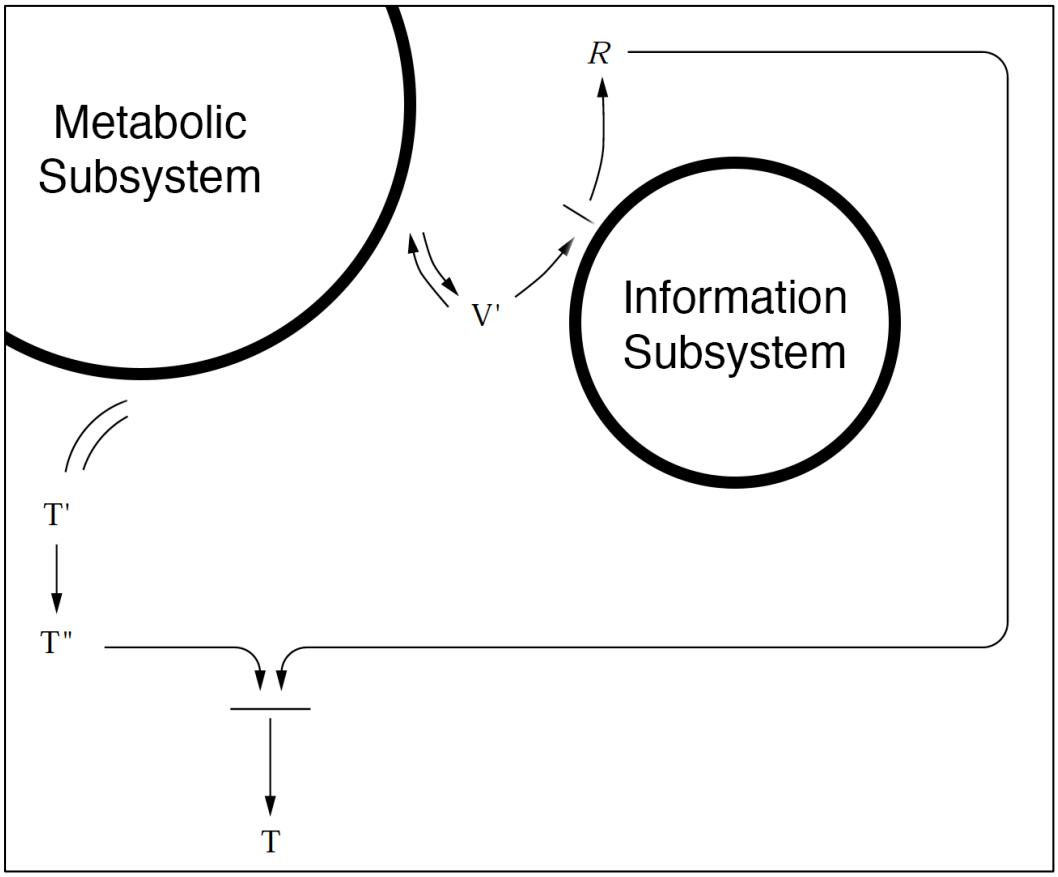


Figure 9.33: The production of T monomers from R and T', in Gánti's chemoton. R is produced by the information subsystem. T' is produced by the metabolic subsystem and spontaneously reacts with something else (perhaps some member of X, or perhaps another T') to form T". Then, R and T" interact (perhaps R acts as a catalyst, or perhaps it is a reactant) to produce T. Figure adapted from Gánti (1997).

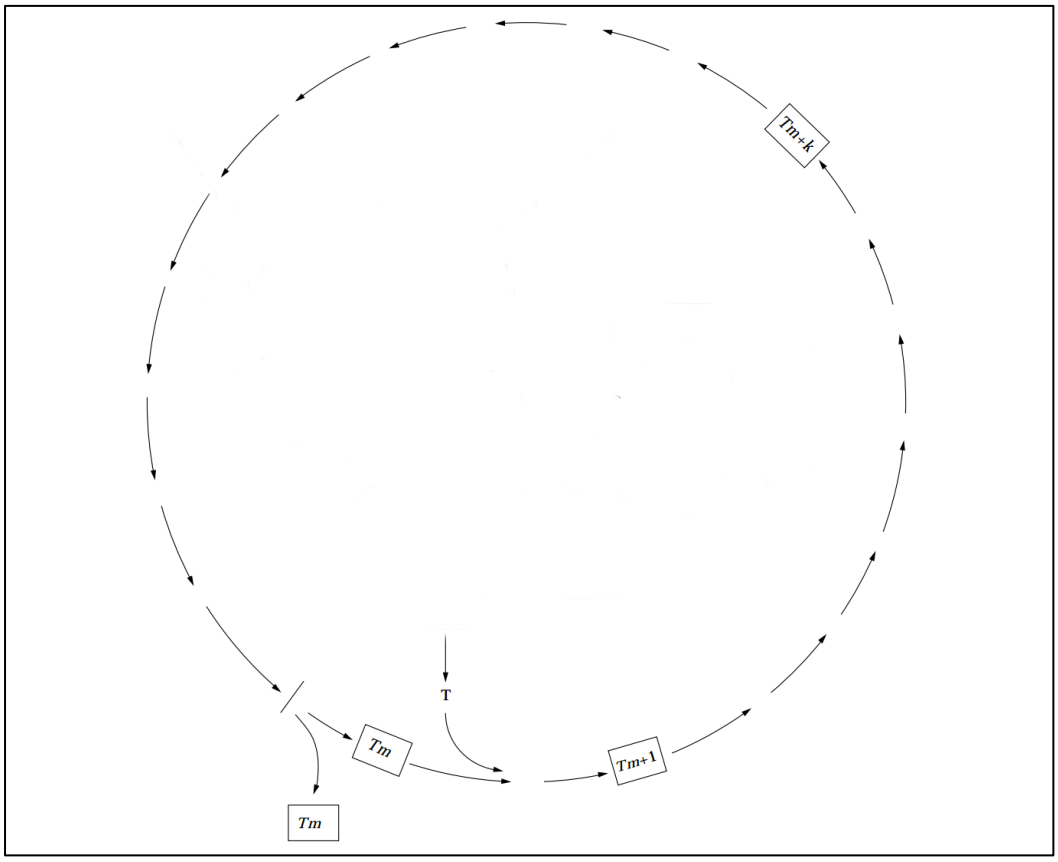


Figure 9.34: The membrane-production subsystem of the chemoton. The membrane polymer, T_m, is composed by the spontaneous assembly of *m* monomers of T. This process is analogous to biological membrane formation, which occurs in much the same way; in the biological case, the monomers are amphiphilic lipid molecules that prefer to align with one another in aqueous environments interspersed with protein-based ion channels (which allow foodstuffs to come in and wastes to go out). The chemoton model assumes that a membrane can be composed of anywhere between *m* and 2*m* monomers but, by the time its size reaches 2*m*, the membrane will divide into two similar membranes (and theoretically, each would contain inside it a complete set of the required members for the other two subsystems). Figure adapted from Gánti (1997).

We can separate the organizational rules for the membrane-production subsystem along the same lines as we just did with the diagrams of these processes. The first two rules, for the synthesis of T monomers, are just the following:

$T' + T' \rightarrow T''$ (or $T' + X \rightarrow T''$)

$R + T'' \rightarrow T$ (or $R + T'' + X \rightarrow T + R$)

The first rule might correspond to the spontaneous combination of two T', or else the spontaneous combination of one T' with one additional molecule (from the set X). In a system where T' and X are freely available, T'' is then spontaneously organizing, meaning it is organizationally present in every state of the system. The second rule might correspond to the spontaneous combination of T'' with R, or it might instead correspond to the combination of T'' with some member of X, catalyzed by R. Again, in an environment where T'' and R (and possibly X) are available as organizational assumptions, the organizational potential of T exists in every state of the system.

The rest of the rules required for membrane synthesis correspond to the serial growth and then division of the membrane as a polymer:

$T_m + T \rightarrow T_{m+1}$

$T_{m+\ldots} + T \rightarrow T_{m+\ldots}$

$T_{m+m-1} + T \rightarrow T_{m+m}$

$T_{m+m} \rightarrow T_m + T_m$

This series of organizational transformations bears some similarity to the set of rules that represented template replication in the information subsystem. Here we have another classic case of autocatalysis, wherein the T_m from the beginning of the series is the catalyst that comes out unscathed at the end, and yet, in that time, it also helps to produce another T_m. The first three rules represent the growth of the membrane to twice its size (from size m to size $m+m$), by way of

sequential addition of T monomers. The last rule represents the division of the enlarged membrane during binary fission (see Figure 8.4). And the only thing that is required to fuel this otherwise spontaneous process is the provision of T monomers, which, as we saw a moment ago, are built spontaneously from fragments earned through the trade with the other subsystems that, themselves, depend on the complete T_m.

Causal Equivalences

To say that a membrane T_m catalyzes the formation of another T_m may, on the surface, appear a bit deceitful because neither of the daughter membranes is likely to have its constituent monomers arranged in precisely the same way as those of the parent membrane. The latter T_m polymers are most likely not the same macromolecules as the former ones; they are just other polymers made of roughly the same number (and type) of monomers.

Similarly, in the case of template replication, different versions of the long polymeric strand (pVn) will bend and twist in slightly different ways, despite containing the same set of monomers strung end-to-end. And so, not only do autocatalytic processes such as membrane fission and template replication seem to stretch what we mean when we say that a catalyst remains unchanged throughout a reaction sequence, but they also seem to stretch what we mean when we say that a pattern is replicating.

The solution to this conundrum is rather simple, however, and it is analogous to the answer we gave earlier when discussing the vibrational variations of smaller molecules and atoms. The many subtle variants of a pattern don't threaten the organizational identicality of systems they are involved in simply because the variants themselves are organizationally identical with one another. That is to say, when coupled with the other subsystems, each variant of the membrane contains the

same organizational potential as the others. And so they form a natural family of variations that are able to cause one another (through their causally equivalent interactions with the other coupled subsystems).

I will use the term "causal equivalence" rather than "functional equivalence" to describe this concept, because what is important to us here is that each of the variations that we will take to be equivalent can cause the same things, although not all of them are necessarily functional. As we have seen, I take the term "functioning" to mean that something serves a goal-directed system, but causal equivalence might sometimes occur in spontaneously organizing systems, not just teleological systems. Two slightly different versions of a bowl, for instance, might be causally equivalent in providing the dynamics for a marble to come to rest at their bottoms, without either of them necessarily being functional; two lumps of platinum might be causally equivalent in catalyzing ethane production without either necessarily being functional (I refer the reader back to Chapters IV and V for an analysis of when I suggest it is and is not appropriate to use the teleologically loaded term "function").

Let's look a little more closely at the case of the membrane to try to understand how we might graph the causal equivalence of various family members. We can specify a series of possible membrane structures that might participate in a chemoton, based on the number of monomers in each version $\{T_m, T_{m+1}, T_{m+2} \ldots T_{m+m}\}$, and we can consider all of these structures to be equivalent in their ability to catalyze other reactions in the coupled subsystems, as long as each of them does so. There is also some other substantial set of similar structures $\{T_{m-1}, T_{m-2}, \ldots, T_{2m+1}, \ldots\}$ that are either too small or too large or otherwise incapable of being causally equivalent in this regard. We can call these sets J and K; J is the possible membranes that work, and K is the possible membranes that don't.

Furthermore, we can consider each of the members of set J to stand in for a more detailed set of variations based on the microstructure of each variant. What I mean by this is that there are many ways that a certain number of monomers might be bonded together, and so the symbol T_m really refers to a number of similar isomers, as do T_{m+1}, T_{m+2}, and so on. And each of those geometries is itself really shorthand for many vibrational, rotational, or stretched variants, as the monomers that make up the polymer—and even the atoms that make up the monomers—constantly wiggle and jiggle. Altogether, there is vast variation. Sets J and K are both really very large.

Focusing now just on set J, let's look at a few of the ways that the many items in the set can be said to be organizationally identical with one another. For one thing, there may be vast subsets of J that, through their wiggling and jiggling, transform into one another and back again. We already looked at this idea with the lock-raking analogy earlier. All of the members of those subsets are, within each subset, organizationally identical. Although they are not catalytic transformations, the differing forms of pattern transform from one to another freely.

For another thing, in the case of the membrane for instance, we understand the spontaneous chemistry of membrane formation to allow the addition and removal of monomers from the polymer, as well as side-slipping of the inter-monomeric bonds within the polymer. All of this results in free interchange between another vast subset of potential membrane states, and if—as we know the case to be—most or all of the members of that subset are causally equivalent in what they do to serve the other parts of the chemoton or cell, then we can consider them to be organizationally identical. Perhaps some members of J are not causally equivalent with the others in this way. So, for those cases, we can just move them to set K.

Lastly, for some kinds of structures, subsets of J that are not necessarily able to inter-transform in the preceding ways, but which possibly form separate not organizationally identical

islands within J, may still have the chance to inter-transform by another method. If some members of two (or more) distinct subsets of J are able to play a causal role in bringing about other structures (say, J.1 → A and J.2 → B), and if those other structures are also able to create members of either (or any) of the distinct subsets of J (so, A + B → J.1 and A + B → J.2), then the subsets of J are able to inter-transform not by transforming into one another directly, but by carrying the potential organization for one another and passing that information into the other structures (A and B) which then may construct a different subset of J. In this case, J.1 and J.2 are organizationally identical and capable of causally and functionally replacing one another, even though they are unable to transform into one another.

There is much more work to be done to formally prove the validity of this system. But for now, I hope it seems reasonable to think that causal equivalences of the sort being described here account for the functional equivalences we see in functional systems, and very likely pose no danger to the notion of organizational identicality.

Putting the Chemoton Together

We can take a closer look at how the chemoton comes to form an autocausal identity by putting together all of its pieces and looking at the entire set of organizational rules that make it up. When we do so, we will find that the system creates all of its own catalytic factors and even some of the foodstuffs that feed individual reactions within it. In order to make that analysis tractable, we can reduce the set of organizational rules over environmentally provided food sources and discarded wastes (X and Y), and then combine piecewise reactions into single steps (by compressing the intermediates). When we have done so, we are left with the following stripped-down set of rules:

$T_m + E + A_1 \rightarrow A_1 + A_1 + V' + T' + N + E + T_m$ Metabolism

$T_m + pVn + E_6 + N \rightarrow pVn + E_6 + R + E + T_m$ Informed Production

$T_m + pV_n + V' \rightarrow pV_n + pV_n + T_m$ Template Replication

$T' + T' + R \rightarrow T$ Membrane Monomer

$T_m + T \rightarrow T_m + T_m$ Membrane Assembly

Analysis of this set of rules shows that each major catalyst in the set {T_m, A_1, pV_n, and E, including E_6} (all showing up in yellow on both sides of one equation or another) and each of the non-environmentally-assumed foodstuffs (V', T', N, R, T) (in blue, on the food side of an equation) is produced within the set (in green, on the product side of an equation), and is not provided by the environment.[332] What that means is that, as a whole, the bundle of three subsystems forms a complex autocatalytic set or hypercycle, the states of which all contain the same organizational potential. The chemoton is a (chemical) system that is able to go on "doing something" ("moving, exchanging material with its environment, and so forth . . .") in the environmental presence only of the members of X. Because of this, we know the organizational graph of the set to be composed of at least one major organizationally identical SCC that encompasses a great majority of the nodes of the organizational graph, and one disjoint minor SCC that encompasses the Ø node of the graph.

The system is organizationally dependent not just on the environment, but upon itself. The set X of various environmental foodstuffs consumed by the chemoton is a necessary set of organizational components, but it is not sufficient, because that set alone does not contain the

[332] *E.g.* A_1 is a catalyst that is produced in the system (eq. 1) and that is involved in producing V', T', N, and itself (eq. 1); E is a set of catalysts that are produced in the system (eq. 2) and that is involved in producing R, A_1, V', T', N and itself (eqs. 1 and 2); T_m is a catalyst that is produced in the system (eq. 5) and that is involved in producing V', T', N, pV_n, R, E, and itself (eqs. 1, 2, 3, and 5); pV_n is a catalyst that is produced in the system (eq. 3) and that is involved in producing R, E, and itself (eqs. 2 and 3); E_6 is a catalyst that is produced in the system (eq. 2) and that is involved in producing R and E, including itself (eq. 2); and V', T', T, R, and N are foodstuffs that are produced in the system (eqs. 1, 2, and 4) and that are involved in producing T_m, pV_n, E, and R (eqs. 2, 3, 4, and 5).

organizational potential that specifies the chemoton. A chemoton is exceedingly unlikely to spontaneously form from an environment of X alone; the bulk of its potential organization is contained in the structures of the three coupled subsystems within itself.

Micelles

In the chemoton model, we saw a process in which a membrane, by way of commerce with two other subsystems, serves as a part in an overall autocatalytic set. It will be useful to compare membranes, which play that integrated functional role in a vitalistic system, with micelles, which are chemically very similar, yet are generally viewed as neither functional nor vitalistic.

One way to think about this issue is in terms of the causes of growth, repair, and replication in the two types of structures. As I noted when introducing micelles, their growth and their "healing", as well as their relatively rare replication[333], occurs only in the same environmental conditions as those in which they begin to form—the spontaneous chemistry of their environment fully accounts for their existence. On the contrary, in the case of cell or chemoton membranes, we must account for their existence not only in terms of environmental conditions but also in terms of the cyclical relationship between the various parts of the cell or chemoton that *create* some of the conditions conducive to their membrane formation. The blueprints for the micelle reside entirely within the environment and not at all within the micelle; if the right parts are available, then micelles will form and if not, then they will not. But the blueprints both for the living cell and for the chemoton reside partly within the environment and partly within that cell or chemoton; there is organizational potential that is requisite to membrane formation hidden within the coupled subsystems of the cell. *Omnis cellula e cellula.*

[333] For example, if a micelle becomes broken in half, the two parts may continue to grow into distinct individual micelles by subsequent addition of further phospholipid monomers.

We can now also describe this contrast in terms of our organizational graphs and rules. In the case of the chemoton membrane, the rules of production require co-catalysis, whereby the membrane enables and ensures the operation of the chemical cycles that produce the monomers that construct the membrane. There is an autocausal loop that forms an SCC that is exclusive of the Ø node of the system. In the case of the micelle, the monomers that spontaneously assemble into the micelle are all provided as environmental assumptions, and so the SCC that would form in that system would be singular, indicating that all states of the system (inclusive of the Ø node) contain the same organizational potential.

This distinction begins to give a hint at why we should not consider patterns such as micelles (or crystals or Bénard cells or storms) to be teleological, alive, vitalistic, subjective, projective, or what-have-you. Each of those patterns may have an identity—a series of states that contain the same organizational potential—but unlike chemotons and other autocatalytic sets, those identities are not self-constructing in the Kantian sense, and, as we'll see when we look next at the concept of value, their resistance to braising (their growth, healing, and replication) is of a qualitatively and quantitatively different nature than that of autopoietic systems.

Autopoiesis

By now we've cast all of the more specific theories of Kantian autocausality that we looked at earlier in terms of our organizational graphs. The only one that we haven't yet looked at is autopoiesis, but that is only because autopoiesis is a more general concept, which subsumes the rest. Autopoiesis, as Maturana and Varela put it, is the self-maintenance of a set of *processes*, and although the systems we have looked at have been specified in terms of the self-maintenance of a set of *structures*, these two formulations can be thought of as being more or less the same, since a structure

or pattern just *is* a set of potential causal proclivities that contribute to processes. Each of the sets of structures (or processes) that we have looked at is more or less an autopoietic or autocausal system that works to retain its own identity (its potential organization) through various processes of self-construction.

I propose that this just is the notion that Maturana and Varela were working to describe in their work. Now, however, we have a way to say more precisely what defines an identity in an autopoietic system, allowing it to truly be the same across time. The notion is generic in the way that I think Maturana and Varela would have liked it to be, but it can also be used to examine specific patterns and the structures and relationships that compose them.

The Draft of the Iceberg

As it turns out, the phenomenon I've just identified as autopoiesis, wherein the members of a set of structures serve as the blueprints for one another, also happens to describe replication. Typically thought of as independent, these phenomena are really two sides of the same coin—both are manifestations of organizational redundancy that actively produces itself. We can try to see this in a couple ways.

Perhaps we can best see the equivalence between autopoiesis and replication in the minimal case of an individually autocatalytic pattern, where the two phenomena are largely convergent. The individual autocatalytic pattern plays a catalytic role in producing more of itself, providing a clear case of replication (as long as the rate of production is higher than the rate of braising; see Lotka 1910).[334] However, if we look carefully, the behavior can also be seen to be identical with that of a

[334] One might imagine some kind of replication in which a replicating molecule is not an autocatalyst, but if we look closely we find that, in those cases where something is produced without being involved in its own production, the product is merely a result of a spontaneous process.

Kauffman-style set of autocatalysts, where the individual parts work to produce more of one another. The only difference is that, in the case of the individual autocatalyst, the term "one another" just refers to the individual itself. As we saw when we first introduced autocatalysis, a basic replicator just is an autocatalytic set of one.

If we look at larger autocatalytic sets now, we find that the picture remains largely the same. When a set of patterns catalyze one another, they are autopoietic (as A may rebuild any lost or damaged B, while B may also rebuild any lost or damaged A) but, when the rate of production exceeds the rate of braising, those parts will be able to produce one another, sufficiently that they will then come to be seen as a replicating set (that is, A and B together create more of A and B).

I will have to leave for future work a deeper analysis of binary fission, but we can see that the general notion of that more complicated type of replication flows out from autopoiesis too: If an autopoietic cell produces its own parts faster than the replacement rate, it will eventually grow too big for its britches, at which point the doubly large cell could split at the seams and become two copies of "the same" cell.

The other way we can see the link between autopoiesis and replication is to look again at our graphs. The graphs that we have been examining up until now can be considered the tip of the iceberg, in terms of the actual organizational possibilities that might describe a system (although our icebergs are upside down because, in this case, the tips are at the bottoms of the graphs). What I mean by this is that a graph in A-B space need not result only in a top node that contains one copy of each pattern (such as node AB). As long as the foodstuffs from which A and B are produced still exist in the system, that top node is also able to produce more A or B, resulting for instance in nodes ABA or ABB. Full graphs of replicating systems would go on and on in an infinite space, expanding beyond AABB, to AAABBB, AAAABBBB, and so on . . . We saw a hint of this already in Figure 9.22, when we first analyzed autocatalysis and the top state of the graph was CC, for a

system that only contained the autocatalyst C (and its parenthesized foodstuffs). But we can have another look now at a system with more parts, in Figure 9.35. As one can see, any graph of autopoiesis specified simply in a space such as A-B-C can be a shorthand representation for a potentially much larger space of replication.

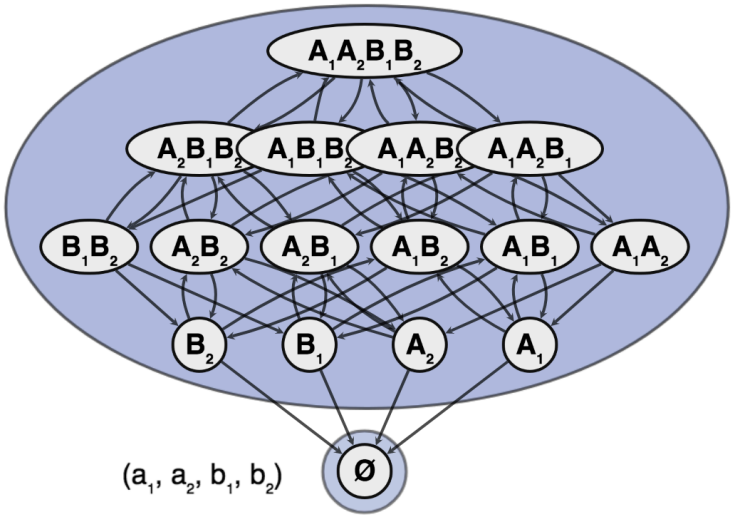


Figure 9.35: An autocatalytic set with two parts, A and B, allowing multiple (for the moment just two) possible copies of each pattern. In this system we see not only how the two parts support one another's existence as an autocatalytic set, but also how they may go on to produce more of one another, potentially resulting in replication of the set of both patterns. This graph could of course be expanded further, to have arbitrarily many copies of both A and B in the top node.

Exploring both autopoiesis and replication in the abstract gives some hints at the generality of our graph-theoretical system . . . but there is a lot more work to be done to show—and prove—just how general this system of classifying patterns might be, and what features of biological

behavior and development might be explainable in its terms. Most of those explorations will have to be put off for future work but, at this early stage of theoretical exploration, one important feature that still needs to be explored is how our mathematics of identity is also able to give us a measure of *value*. As we'll see, the graph-theoretic structures we have been working with can be used to measure just how an autopoietic identity can, through its coordinated internal machinations, persist longer than it otherwise would.

Chapter X

Value

The day that the universe contained entities that could take rudimentary steps toward defending their own interests was the day that interests were born.

—Daniel Dennett (1984, p. 22)

All sciences now must do the preparatory work for the future task of the philosopher: understanding this task to be, that the philosopher has to solve the problem of value.

—Friedrich Nietzsche (1887)

One of the most striking facets of Richard Dawkins' gene's-eye-view account of natural selection is that it is given—and can only be understood—in terms of *benefit*; it is based on the idea that genes may have a personal, evaluative perspective on the world. Genes and their behaviors are seen as being *selfish* and it is their *good* that the activities of their bodies-as-vehicles are said to serve (*cf.* Dawkins 1976, 1982). As we saw in Chapter II, this benefit-based view of biology faces a serious challenge when confronted with materialist sensibilities whereby molecules such as DNA are objective patterns, forbidden from having subjective properties.

And yet it is unavoidable. While the champions of the gene's-eye-view have tended to back off from literal claims of individual agentive molecules that actually benefit, they have been unable to relinquish the value-laden rhetoric of selfishness, and the prosecutor's legalese that Dennett (1995) has appropriated into biophilosophical inquiry: *cui bono?* Who stands to benefit?

Furthermore, while these gene's-eye-view theorists openly acknowledge the notion of benefit, judgments of value are by no means unique to their perspective. Competing accounts of the units of selection are all framed in terms of some kind of persisting unit (cells, organisms, species, kinship groups, developmental systems). And no matter which of these persistors we embrace, we understand a great many of their attributes and performances as contributing to—that is, as *being good for*—their persistence. That might sound controversial to some anti-adaptationist thinkers, but it shouldn't: the "for" in this "good for" does not need to imply "meant for", "designed for", "intended for", "adapted for", or anything of that ilk. We can leave competition, adaptation, evolution, and design aside entirely and still say, merely by definition, that the things that contribute to persistence are "good for" persistence . . . and thus good for the persisting pattern's very existence in the world.

The notion of benefit crops up independently in another realm of biology, too. The phenomenon of autopoietic survival is an alternate kind of persistence that is just as difficult to characterize without notions and rhetoric that are, at heart, evaluative. We are unable to speak of food, water, or other resources used by autopoietic agents without implicitly conceding that these things are *beneficial* to those agents; likewise, there is no discussion of disease, predators, or other dangers that doesn't at least tacitly admit that these things are damaging to the agent or organism. It is both tautological and hackneyed to repeat at this point, but I feel I must: in a universe where any particular pattern's existence is far from guaranteed, the pattern's continuing to exist is good for the pattern's existence, and its not doing so is bad. At least one kind of value—at least one very central sense of the terms "good" and "bad"—derives directly from the existential notion of persistence.[335]

So here is my theoretical offering: I suggest that the objective property of the natural, material world that underlies the subjective properties of value and benefit is simply time. An

[335] I am convinced that this is the only sense of value or benefit that matters, and all other senses of those terms derive from this. But for now any argument for that position will have to be postponed.

identity benefits when it is granted more time. A thing is valuable (to a persisting identity) when it is able to grant that identity more time.

Using the notion of organizational identicality that we've just developed, I'm going to argue that the foundation upon which value is built is the amount of time the potential organization within an identity might be expected to continue existing. The context for persistence is of course that of ratcheted environmental braising. And in that context, we will find that there are some organizational identities that, if they do come to exist, are not long for this world, but there are also some organizational identities that have methods by which to actively resist the otherwise statistically assured decay. Despite the shortcoming that our measurements of time here are given in ticks rather than physical units such as seconds, we can nonetheless abstractly quantify the differences in how long these different kinds of identities persist.[336]

[336] Perhaps one day a more physical model could be built from these abstractions if we were to make some adjustments that probably would include, at the least, altering the Markov chain to a continuous-time version by specifying a transition-time matrix specified in units proportional to seconds.

A. Measuring Persistence

Differential persistence takes a variety of forms. Some patterns persist only as long as they do not encounter other patterns. Others persist through some interactions, while undergoing dissolution or transformation in others. Still other persistent patterns interact with only a few other patterns, simply maintaining their form in all other contexts.

—John Holland (1998, p. 227)

I am going to offer what I think to be the most straightforward ways of measuring the amount of time that an identity might persist. There are a couple of ways that those measures will be imprecise, but the lack of precision is no reason for concern. First, the measures turn out to be probabilistic (because of the probabilistic nature of our Markov processes). This is good news, since probabilistic measures reflect the uncertainty of the real world much better than precise measures. Second, because an identity is made up of many different possible states, the amount of time such an identity may persist must be calculated independently for each of those states as a possible starting state. Every node in an identity is of course organizationally identical, but still, an unhealthy identity—where the current state contains less redundancy--cannot be expected to last as long as a healthier one.

As our foregoing analysis suggests, the organizational potential in an identity continues to exist as long as the current state of the system contains the potential organization for all of the states in the identity. And that remains true as long as the state of the system is a member of the SCC that outlines the identity. Our metric for persistence, then, needs only measure how long it will take before the state of the system exits the SCC.

Actually, I have so far found two metrics to be of interest in this regard, although there may be other possibilities. The two are what I will call the *expected lifetime* and the *relative expected lifetime* for each node within an organizational identity. The expected lifetime is just a standard statistical expectation of the number of time steps from a particular node within the Markov process to any node outside the SCC, and calculation of that expected lifetime for any node results in a number of ticks. The *relative* expected lifetime is the ratio of that expected lifetime to the baseline lifetime of that node, where baseline lifetime is, as we'll come to see, related to how long the patterns in that node would resist braising by their resilience alone. The relative expected lifetime is a unit-free number that tells how many times longer the potential organization of the structures in the node is likely to persist when the node is a member of this SCC than when the patterns in that node are not a member of the SCC. We'll piece this all together now to make it clearer.

Baseline Lifetimes

The notion of persistence would be of no interest if patterns were either so resilient that everything lasted forever or so fragile that nothing lasted more than a moment. It is of interest, however, because there is a baseline amount of time—less than forever and more than momentarily—that a set of patterns might naturally last and, more particularly, because that amount of time can potentially be changed.

One way to think about the baseline lifetime for a set of patterns is in terms of how long the actual organization in those patterns would continue to exist, given only the braising effects of the current environment and ignoring any constructive activity that might rebuild the patterns.[337] In order to calculate this value for each of the nodes in a graph, we first need to simplify the graph so

[337] In other words, how long it would last if the state's *potential* organization were limited to only its *actual* organization.

that it reflects only the down-bound transitions of braising, and none of the up-bound transitions that correspond to organizational efforts. Once that is done, the baseline lifetime for each node is just the expected number of time steps before the process exits that individual node.

The simplest diagram displaying unchecked braising of this sort would consist of a node for one pattern, and then one downward transition from there to a Ø node (representing the absence of the pattern). Such a graph denotes a pattern left to fall apart in an environment where there exists no potential to rebuild it. We saw an example of this earlier in Figure 9.8, which, at the time, I said might represent a coconut tree that had some risk of being knocked over, but no chance of returning to a standing position once it is downed. We could also think of it as a biological macromolecule such as a strand of DNA outside of a cell, which will eventually decay by radiation or chemical attack, and will not get rebuilt. Because there is no up-bound return transition, the two nodes in this kind of graph are not organizationally identical to one another; each one forms an SCC of its own. The expected lifetime and the baseline lifetime for node A in this case are the same—each is the reciprocal of the probability p that node A would transition to the Ø node. So if p is 0.01 (1%), then the baseline lifetime here—the expected amount of time before the transition would be taken—is 100 time steps.

I think that image is clear enough, but we can gain a bit more insight into the calculation by looking at a graph of a scenario with a few more parts. In Figure 10.1, the patterns S, T, and U will all tend to disintegrate over time. Nothing in this environment plays a role in the formation of any of these patterns.[338] As we can see in the figure, each node of the graph is an individual identity; none is mutually reachable from any of the others. Therefore, every change of state in this graph is an irreversibly destructive transition to a different organizational identity.

[338] We might imagine that S, T, and U actually form an autocatalytic set in which case their expected lifetimes will be more interesting, but in order to calculate the baseline lifetimes, we must use a simplified graph that ignores any constructive capacities.

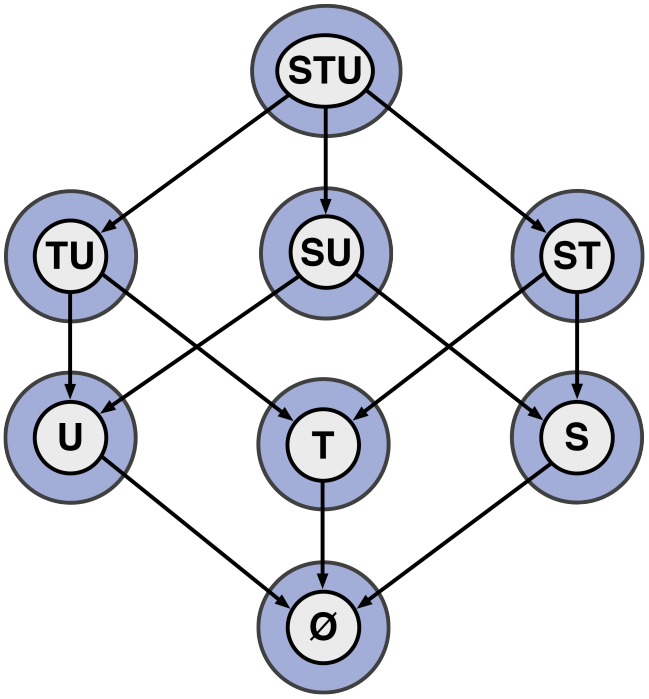


Figure 10.1: An organizational graph of patterns S, T, and U. In an environment in which we have introduced some S, T, and U, but none of these patterns has any tendency to form (either spontaneously or catalytically), we will see unfettered braising of all three patterns. This will eventually drive the system by one of six possible paths to the null state, where none of the three patterns exist. Every transition in this graph is irreversible, and so no state is organizationally identical to any of the others.

In graphs like this, which have no cycles and in which every node is its own SCC, the lifetime of any particular node turns out to just be the reciprocal of the sum of the node's outbound transition probabilities.

For instance, in the Ø node, there are no outbound connections and so, once the system arrives there it will stay indefinitely—the lifetime is the reciprocal of zero, which is infinity. To calculate the baseline lifetime for any of the other nodes, we'll need to specify transition probabilities. To keep things simple, let's just assume that all the connections in this graph have a transition probability of 0.01. That is to say, there is a one percent chance that, with any tick of the

clock, energetic events in the environment might damage any of S, T, or U, causing a transition to another state. In any of the single-letter states (the row above Ø), there is precisely one outbound connection, and so the amount of time we can expect the system to remain in that state will be a hundred time steps (one divided by 0.01). Each of the identities in the next level up (ST, SU, and TU) has two outbound connections with transition probabilities of 0.01, and so any of those nodes can be expected to decay in roughly fifty time steps (the reciprocal of 2×0.01). And the topmost node, STU, has three outbound paths. The reciprocal of 3×0.01 tells us that that node has a baseline lifetime of roughly thirty-three time steps.

Expected Lifetimes

Some systems, or some patterns within those systems, may actually have no constructive causal powers within them, and so the expected lifetimes of the nodes in those systems won't differ from their baseline lifetimes. As with the example above, they will just be subject to a slow but inevitable decay due to the distribution of braising energy around them.

But when a graph contains SCCs that are richer than just one node, there exist cycles in those SCCs that allow the state of a system to potentially circulate, for a while, before eventually decaying downward. The expected lifetimes of nodes in these SCCs are thus potentially greater than the reciprocal of the sum of their down-bound transitions. We still can determine those expected lifetimes by extending our calculations using a bit of linear algebra. For those readers familiar with Markov processes and matrix math, the calculation we're going to derive now will be fairly straightforward. For those not inclined to follow along with the upcoming formulae, it will be sufficient to attend only to the conceptual-level description of what those equations mean, and to pick the thread up again about six pages later.

In order to simplify our task, the first thing we can do is to separate the strongly connected component that interests us from the remainder of the graph. This requires the fabrication of what we can call a "sink" or an absorbing state—a node to which any outbound connection from any of the nodes in the strongly connected component can be redirected, maintaining those connections' original transition probabilities.[339]

So, for instance, if we are looking at the diagram in Figure 10.2.a, in the space of patterns A, B, and C, and if our interest is in the strongly connected component highlighted in beige wherein A catalyzes the formation of B and B catalyzes the formation of C, then we would prune that graph to just that four-node SCC, with a sink node added to make a five-node graph, as in Figure 10.2.b.

[339] If a node has multiple outbound connections to different states outside the SCC, the transition probability of the new singular connection from that node to the sink state will simply be the sum of the previous connections' transition probabilities.

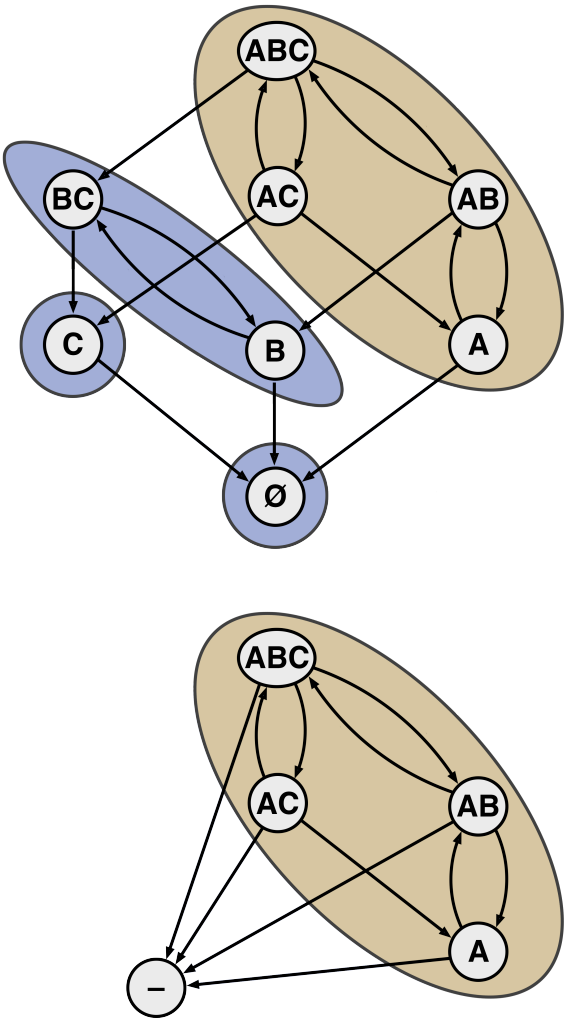


Figure 10.2: The process of extracting an SCC from a more detailed graph in order to facilitate the calculation of expected lifetimes. a) A graph wherein A → A + B and B → B + C. The relevant SCC in A-B-C space is highlighted in beige. b) The same strongly connected component has been divorced from the remainder of the graph, with a single sink node added to absorb the SCC's outbound transitions. Another way to think of producing the diagram in (b) from the one in (a) is to simply consolidate the nodes in the remainder of the diagram until they become one. The transitions out of the SCC into that consolidated node retain both their source nodes and their transition probabilities. The key here is simply that these transitions represent the ways that the current state of the Markov process might exit the SCC, and thus they represent the moments in which the state of a system loses its organizational identicality.

Once we have this simplified graph, we can derive the expected time to exit to the sink state from any state in the SCC, by setting up a recursive sum for each node and then solving the set of these sums as a system of linear equations (see Grinstead and Snell 2006, pp. 419–420). In short, the expected lifetime for an SCC, beginning at any particular node, is equal to one (the number of steps it takes to transition out from the current node) plus the sum of the expected lifetimes beginning at each of the other nodes, individually weighted by the probability that the system will wind up in any of those nodes during the transition out from the current node.

We can look at a couple of simple examples of this to get an intuition for what it means. For instance, if, in a very simple graph, there is a 100% chance of transitioning from node *i* to the sink state, and if the expected lifetime of the sink state is zero (that is to say, once you're in the sink state, you are already out of the SCC and so there is no lifetime for that SCC), then we end up with:

$$E[L_i] = 1 + (1 \times 0) = 1$$

In this example, I am using the probability theorist's notation $E[L_i]$ to denote the expected value, $E[\ ]$, of the lifetime, L, of the SCC, beginning at the particular node, i, in the graph. The expected lifetime in this case is a total of one time step, because our first time step (our next transition) is guaranteed to be a departure from the SCC.

If our system instead has a 50% chance of self-transitioning to (*i.e.*, remaining in) the current node and a 50% chance of transitioning from the current node to the sink node, then we have:

$$E[L_i] = 1 + (0.5 \times E[L_i]) + (0.5 \times 0) = 2$$

The expected lifetime is now two time steps. Sometimes the lifetime will be less and sometimes it will be more, but on average, it will take two time steps before the system transitions to the sink node, and from there it will never return to the SCC that contains node *i*.

In the general case, we can assume that there are *n* nodes in the graph (excepting the fabricated sink state, for reasons that will become clear shortly), and we can denote the probability *p* of transitioning from node *i* to node *j* by $p_{i,j}$. The verbose formula for the lifetime of the SCC, beginning from node *i*, can then be written as:

$$E[L_i] = 1 + \left(\sum_{j=1}^{n} p_{i,j} \times E\left[L_j\right] \right) + \left(p_{i,sink} \times E[L_{sink}]\right)$$

The first term here—the "1"—represents the cost of the first transition from node *i*. The next term—the sum—represents all of the *n* potential future lifetimes that might result from the current transition landing in any of the *n* nodes of the SCC. The final term similarly represents the potential future lifetime, given the $p_{i,sink}$ chance of arrival instead at the sink node (any node of the original graph outside the SCC). This last term is of the same form as the summed terms and so it could easily have been included as an n+1st iteration of the summation; however, since it will always evaluate to zero (because $E[L_{sink}]$ is zero), it is better left out of the sum, and can be struck from the overall equation entirely, leaving us with the following simpler formula for the lifetimes of any of the *n* nodes in an SCC:

$$E[L_i] = 1 + \sum_{j=1}^{n} p_{i,j} \times E\left[L_j\right]$$

Since each $E[L_i]$ depends on the answers to all the other $E[L_j]$, in all but the simplest cases we must turn to linear algebra in order to solve the set of them, all together, as a system of equations.

One way to do this is to move all the unknown variables to one side of the equation,

$$E[L_i] - \sum_{j=1}^{n} p_{i,j} \times E\left[L_j\right] = 1$$

and then to duplicate the equation n times to account for the n nodes whose expected lifetimes we would like to determine, and then recast the set of equations in matrix form and simplify. The result of those operations is:

$$(I - P)\, EL = 1$$

where I is an n x n identity matrix; P is an n x n matrix of the various transition probabilities, $p_{i,j}$, in the excerpted graph; EL is a length-n vector of the various $E[L_i]$ values that we want to find; and **1** represents a length-n column vector of ones.[340]

So, since I and **1** are constants, in order to find the expected lifetimes of all the nodes in an SCC, we only need to plug into this formula the matrix P (which is the Markov transition matrix for the truncated graph, with the row and column corresponding to the sink state trimmed out), and

[340] It helps, in understanding this derivation, if one notes two simplifications. First, we can see that the first term in each of the individual equations, $E[L_i]$, is equivalent to the product of the column vector EL and the ith row of the identity matrix (a product that simply isolates the ith value in vector EL). Second, we can note that the weighted sum in each copy of the equation is equivalent to the same column vector EL times the ith row of the Markov transition matrix. Once we see these two equivalences, it makes sense that the entire left-hand side of the equation is simply the product of EL and the difference of the respective ith rows of the identity matrix and the Markov transition matrix.

then to solve the linear system. The result will be the vector EL, which then contains the various $E[L_i]$ that we are interested in.[341]

For those readers who were not inclined to follow along with the math, this solution tells us the *expected* length of time before any particular state of an organizationally-identical set of states decays to a state that is no longer organizationally identical with what once existed.

Those expected lifetimes vary across the different starting states of the system. However, we'll find that the variation is often qualitatively constrained to certain quantitative categories, and so our expected lifetime calculations turn out to be useful in making generalizations about which identities in which systems might be expected to last longer than the patterns in them would last in the absence of any reconstructive activities.

Longer than Otherwise

It is difficult to argue with a child who claims that their act of jumping off the sofa and falling for a moment before hitting the ground is *flying*. And the argument isn't difficult just because you don't want to spoil the child's fun; it's also philosophically problematic because . . . after all, how *do* you define flying? Is it not just passing through the air for a while, as the child seems to think? Wouldn't we say that an intercontinental ballistic missile is flying halfway around the Earth even after its launch thrusters have spent their fuel, or that a baseball is flying through the air on its way out of the ballpark?

Blurry cases such as the suborbital ICBM and a fly ball at a baseball game may make it difficult for us to reach universal agreement about what the word "flying" should mean. But the candidate distinction that most people's minds latch onto to help distinguish between ballistic

[341] Thanks to Eric Nichols for helping me find this analytical solution.

children and airplanes is the difference between powered and unpowered flight. We might tell the child that to *really* be flying, the way birds, planes and superheroes do, they would have to remain in the air longer than they otherwise would and, in this case, we tell the child that we can quantify "otherwise" by saying they would have to be doing more than just falling. Powered flight is not just being up in the air, but involves additional work to exceed the baseline behavior of simply falling, as dictated by gravitation and one's existing momentum. There is a clearly measurable amount of time that one would stay aloft if one had no special properties. To be flying means one can, in theory, beat that benchmark.[342]

The theoretical definition of relative expected lifetime that I am proposing is much like this definition of flying. When a system's organizational content has a lifetime that exceeds the baseline dictated by braising—when, through some kind of activity, the potential organization in a set of structures is able to persist *longer than it otherwise would*—then, I claim, we are justified in saying that that identity is benefiting. It is doing more than just the organizational analogue of falling. It is persisting.

We've seen that a baseline lifetime for organizational patterns can be defined in terms of the expected lifetimes of those same patterns in the absence of their organizationally creative capacities. And we've seen that, because of the cycles within SCCs, expected lifetimes for nodes involved in SCCS can exceed baseline lifetimes. One way to quantify the idea of "longer than otherwise", then, is to use a ratio to compare a node's expected lifetime against its baseline expected lifetime. In other words, we can set up a quotient with the numerator equal to the expected lifetime of the node when it participates in the SCC, and the denominator equal to the expected lifetime of the node when it

[342] Of course there is the exception case of powered flight directed towards the ground, in which case impact may be even sooner than otherwise dictated by gravitational influence. In the case of teleological systems, this correlates with the idea of a thing that destroys itself faster than its individual parts naturally would be destroyed. That might be possible, in theory, but just as with flying, accelerating under the baseline rather than over it is, on one hand, still distinguishable as qualitatively different but, on the other hand, far less interesting.

stands alone as an SCC of its own. The formula below expresses this simple relative expected lifetime ratio.

$$R[L_i] = \frac{E[L_i]}{B[L_i]}$$

We can take as an example the graph we looked at earlier in Figure 10.1. As we already determined, if we assume all the transitions to have a 0.01 probability, the baseline lifetimes of the nodes are 33.33 for the top node, 50 for each of the three nodes below the top node, 100 for each of the three nodes above the bottom node, and infinity for the bottom node. These values can be used as the denominators for determining the relative expected lifetimes, when these nodes participate in various kinds of other SCCs.

So, for instance, if there are no constructive transitions in this system, then the system is as it appears in Figure 10.1, and the expected lifetime of every node is the same as the baseline lifetime, and the relative expected lifetime of almost every node in the graph is 1.[343] What this means is that, starting from any of these nodes, the organizational content of the SCC is bound to decay at precisely the rate dictated by braising.

In general, if the relative expected lifetime of a node in an SCC is less than or equal to 1, then the pattern represented there is a transient part of our world; it is some kind of ontological nonce with no power to persist. If, on the other hand, the relative expected lifetime is greater than 1, then the organizational pattern can be considered to be some kind of persistor. There is something about the pattern or its relationship to its environment that ensures that the potential organization of this identity will stick around at least a little longer than it otherwise would. And

[343] There is an exception for the null node, the relative expected lifetime of which, strictly speaking, is mathematically *undefined* (it is infinity divided by infinity). That seems just fine, since the question of how long no pattern at all might continue to exist is itself best thought of as being undefined.

that is certainly *good* for that identity. In the next chapter, we can begin to use our expected lifetimes and relative expected lifetimes to look at the behaviors of differing kinds of persisting identities.

Chapter XI

Patterns in Time

Like the standing wave in front of a rock in a fast-moving stream, a city is a pattern in time. No single constituent remains in place but the city persists.

—John Holland (1995)

Schrödinger (1944) recognized that in order for the organizing processes of life to get off the ground they must somehow resist the second law of thermodynamics. In his work that began roughly two decades later, Prigogine showed how to do this. It is only within so-called *dissipative*, far-from-equilibrium, open thermodynamic systems that there exists a reliable flux of energetic order that is able to provide the free energy required to do the work to shuffle around any material that is to become materially ordered. This showed how we can, in principle, get order out of chaos.

While Prigogine's paradigm gives us the thermodynamic answers we need, it has not yet addressed the material coordination problem. It still doesn't tell us what *kinds* of order we can get out of chaos, or where in the natural world the information that guides the construction of that orderliness might reside. But the analytical technique we have just developed can help with this. Let's review the details again to see how.

We can begin with the central concern, which is the following: In a world where all that exists is physical patterns, the blueprints and the machinery required to construct any physical patterns must also be physical patterns. And because every physical pattern is subject to the statistically assured decay of ratcheted braising, there is a constant risk of organizational erasure—not just of patterns but also of the patterns that *generate* those patterns (and the patterns that generate

those patterns, and so on).

In order to begin solving this problem, we based our system of graphs around the distinction between actual and potential organization. Since an actual pattern may contain the potential organization for other patterns, each overall graph (of the possible transformations among a set of causally related patterns) reflects the distribution of organization within a system. With that in mind, we can analyze a variety of graphs, in conjunction with the lifetime calculations that emerge from them, to discover the modes of organization that are able to contribute to the persistence of identities. That is, we can look at various organizational identities within graphs to attempt to discover the kinds of organizational relationships between a system's parts that result in various types of lifetimes.

Doing such an analysis will put us in a position to outline more clearly what I would like to call the tripartite ontology of our world—a short, high-level catalogue of the kinds of organization that might exist. There are two points of interest here. Firstly, it is interesting that there exists such a concise catalogue that can be used to categorize our world, as Herbert Simon intended, in terms of "the persistence of the stablest". Secondly, it is interesting that one of the categories in that catalogue serves to outline patterns that are able to help themselves, and thereby to differentiate those patterns from any others. These are the teleological patterns that show just what kind of organization matters for vitality to emerge in a purely material world.

A. Ontological Nonce

Things fall apart; the center cannot hold; Mere anarchy is loosed upon the world . . .

—William Butler Yeats (1921)

Some organizational identities just don't persist. There are no dependable, complete blueprints for them anywhere in their environment. If and when these sets of patterns come to exist, they generally tend to decay at a predictable rate and, because they lack complete blueprints, there don't exist any processes that could fully rebuild them. Identities of this kind are what I referred to in Chapter II as ontological nonce.

Standard Nonce

We have already looked at organizational graphs that represent systems like this. The clearest examples so far may be the graphs we saw in Figures 9.8 and 10.1, both of which represent systems of unfettered braising. If, by random chance, patterns S, T, and U all come to be together in some environment, but none of them has any causal effect in the production of the others, and if none of them is produced by any spontaneous processes in the same environment, then the current state of that system exists as an identity (an SCC) consisting of just a single node—specifically, in the case of S, T, and U, we have the top node of the graph in Figure 10.1.

The behavior of an isolated identity of this sort is only temporary resilience, and then, eventually, decay by means of one of the one-way transitions to a node at a level beneath the

singleton SCC. That initial set of patterns will have lost some actual organizational content while not gaining any potential organization. The loss is irrecoverable.

In ontological nonce of this sort, there is no state elsewhere in the system that has a transition up to a node such as STU. There are no blueprints for S, T, and U anywhere in the system.

The fact that none of the patterns in such a node carries the organizational potential for any of the others means that none of them will persist any longer than braising will allow. This is easily quantified. As we already saw when we were discussing baseline lifetimes, any single-node SCC will have an expected lifetime equal to the reciprocal of the sum of its outbound transition probabilities, and a relative lifetime equal to one (except for the Ø node, which has a relative lifetime that is undefined; see footnote 343, p. 573). What this means is that the expected lifetime of standard ontological nonce is just the baseline lifetime, and that the sets of patterns that make up such states are able to accrue only random and unreliable existential benefit in this environment. They are not meant to be.

Catalytic Nonce

Standard ontological nonce, as we've just looked at it, occurs when *no* pattern in the system is able to help any other pattern. In a short while, we will compare this with Kantian systems where *all* the patterns in the system are able to help one another; however, there are various middle grounds between these extremes. Consider first a system in which we have *some* catalytic causal capacity (*e.g.*, A→AB), but there is no Kantian causal circularity (*e.g.*, there is no B→BA). In such a case, *some* of the patterns are able to help *some* of the others. Figure 11.1 shows a sample graph for a system of this sort.

Again, in a case like this, there is no state of the system that produces all the organizational potential in the identity. For example, in the large SCC in the figure, we find that A produces B, which produces C, but nothing here produces A, either spontaneously or catalytically.

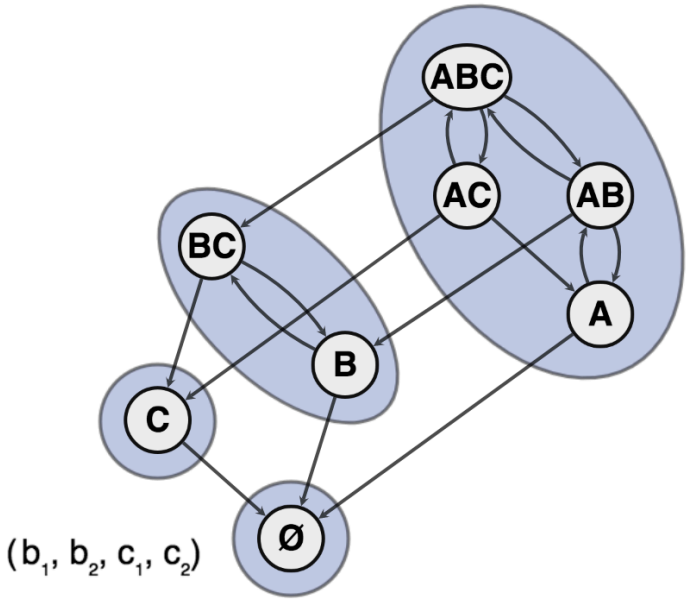


Figure 11.1: An organizational graph of a system in an A-B-C space, where A catalyzes the formation of B and B catalyzes the formation of C. In this system, four SCCs arise: The topmost SCC shows that the existence of A alone is organizationally identical with any state that contains A along with any combination of B and C. Since the potential for both B and C exists within A, braising in the B and C dimensions is nondestructive in this SCC, and so the expected lifetimes of any of these nodes are tied only to the braising rate of A. In the absence of A and in the presence of B, the next SCC down has two organizationally identical nodes whose decay, according to the same logic, is tied only to the braising rate of B. And in the absence of both A and B, the C node tends to decay at its own baseline rate. Overall, the system seems to decay in all directions, but, in the presence of B there can be some temporary but not very significant preservation of the organizational potential for C, and in the presence of A there can be some temporary but not very significant preservation of the organizational potential for B and C.

We can look at the expected and relative lifetimes of the nodes in the catalytic system in the figure. When we do so, we find that all the nodes in the largest SCC have an expected lifetime equal to the reciprocal of the downward transition probability for A-decay, all of the nodes in the second largest SCC have an expected lifetime equal to the reciprocal of the downward transition probability for B-decay, and the C node (in an SCC of its own) has an expected lifetime equal to the reciprocal of its own downward transition probability. And the Ø node, in which neither A, B, nor C exists, has an infinite expected lifetime.

The relative lifetimes of all these nodes are all different, but the bottommost node in each non-null SCC (the A, the B, and the C nodes), has a relative lifetime of 1, signifying that the potential organization in these nodes cannot be expected to survive any longer than it would if the system of these patterns didn't have its catalytic capacities. Higher nodes (AB, AC, BC, ABC) all have higher relative lifetimes because, in the absence of catalytic capacities, these would have been susceptible to braising in multiple dimensions—that is to say, the Bs and Cs in those nodes would be more vulnerable because they are not reparable; however, in this system, the lifetimes of those nodes are somewhat extended thanks to the catalytic capacities of A to help form B, and of B to help form C. One way we might think of this is to say that, for instance, in the largest SCC here, B and C last longer than they otherwise would, *given* A. But since A is *not* necessarily given (because it has its own vulnerabilities), the entire identity—the sum of the potential information for A, B, and C—has an expected lifetime that is dependent solely on the braising rate of A. Because of this, a mere catalytic system of this sort is still unable to persist any longer than the catalyst alone would otherwise persist, and so we must categorize the system as another kind of ontological nonce that decays at the standard braising rate of one of its parts.

Flat Cyclical Nonce

For the sake of inclusiveness, we can analyze one more subcategory of nonce (there may be still others, but for now we will limit our explorations here). In the case of what we might call flat cyclical systems, the patterns in a certain set are able to transform into one another in a cycle, but are not able to help one another except at the cost of losing their own organization. This is what happens in ordinary reversible chemical reactions, such as those in Figures 9.24 and 9.25, but we can look at a more abstract version of the phenomenon now. For instance, pattern A may *become* B, and B may similarly become C, which then may become A, all without *catalyzing* one another's formation. Again, however, in SCCs of this sort, there is nothing either spontaneous or catalytic that *produces* any of A, B, or C. The SCCs that represent this kind of relationship remain at a single level in an organizational graph.

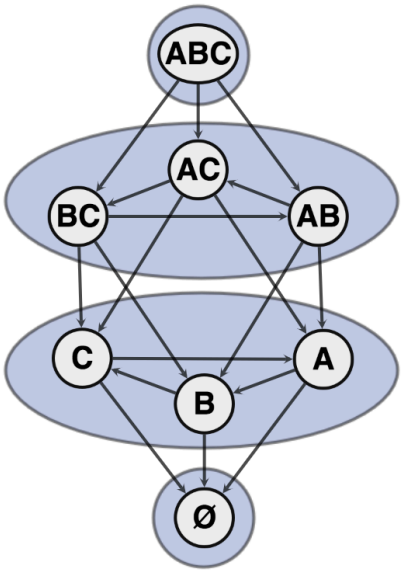


Figure 11.2: An organizational graph representing a system in which A transforms into B, B transforms into C, and C transforms into A, but none of them catalyzes the formation of the others. As we saw earlier, this may be chemically very unlikely, but the case of reversible reactions, in which, for instance, A + B converts to C + D and C + D also converts back to A + B, is extremely common. In any case, all of the nodes in a single-level SCC are organizationally identical with all of the others. But precisely because the SCCs formed by these kinds of causal relationships do not span levels, no redundancy is ever gained through transitions in systems such as this.

When we look at the expected lifetimes for all the nodes in any single-level SCC, we find that they are all equal to the reciprocal of the sum of the probabilities of the node's down-bound transitions—a number that is always equal to the baseline lifetime for that node. And because of this, the relative lifetimes of all such nodes are—just as we saw with standard ontological nonce—equal to 1 (again, except in the case of the Ø node). Again, these limitations constitute a form of ontological nonce in which the combined organizational potential of any set of these patterns is unable to persist longer than the expected amount of time that their individual parts would persist.

Causalytic Poisoning

There is another potential case in which a set of patterns is unable to persist any longer than braising would allow, but in this form of nonce—call it *causalytic poisoning*—the lifetimes may be even shorter than those of the same patterns left to decay by random environmental braising effects. These are the cases in which a pattern within the system actively plays a causal role in the *destruction* of another pattern. So if, for instance, pattern C acts as an enzyme in the catalytic decomposition of B, then there will be an additional likelihood of B decomposing at any time, beyond the braising effects of the environment. In that case, while the baseline lifetime for any node containing both a B and a C will be the reciprocal of the down-bound braising transitions, the expected lifetime for any such node will be less, because it will be the reciprocal of the braising transitions plus the down-bound enzymatic decomposition transition (it will be easier to model with two distinct arrows going between the nodes, representing the different causes of decomposition). With an expected lifetime lower than the braising lifetime, the relative lifetime of such nodes will then be less than 1.

B. Spontaneously Organizing Systems

That crystalline structures and the structures of living beings could be related by applying [the] criterion [of apparent self-construction] . . . might well give [an investigator] food for thought. Even if unversed in modern biology, [the investigator] would wonder whether the internal forces which give living beings their macroscopic structure might be of the same nature as the microscopic interactions responsible for crystalline morphologies . . .

—Jacques Monod (1971, p. 22)

The model of value that we have developed shows that there are some sets of patterns that, because of the causal relationships among them, are regularly rebuilt and thus are able to persist longer than they otherwise would. There is, however, a further distinction to be made between different *manners* of persistence. Some sets of patterns benefit from *external* processes that give them more time and they are, therefore, spontaneously organizing in the presence of those external sources of information. Others have their own *internal* processes by which they buy themselves more time, and it is these internally informed, organizing processes that can be said to be *goal-directed*.

We have already become familiar with graphs of both of these categories of system. In Chapter IX, we saw some sets of patterns that have blueprints that seem to be (metaphorically) written in the stars, and others, analogous to a multiquine, that have blueprints that appear to be (literally) written in themselves. What we can do now is to take a look at the expected and relative lifetimes for the nodes in the SCCs that emerge from those graphs. We will begin here with spontaneous systems.

Standard Spontaneous Systems

As we have seen, in a fully spontaneous system, everything is created from Ø, and there is one SCC that encompasses all the nodes of the graph.

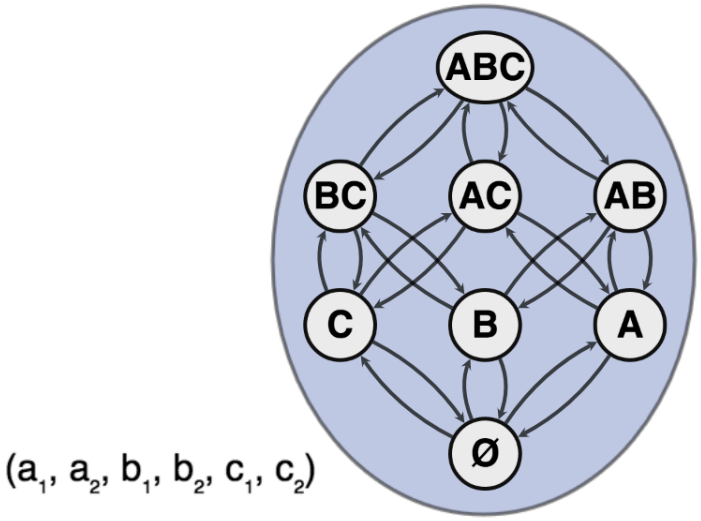


Figure 11.3: An organizational graph of a non-reduced, fully spontaneous system wherein A, B, and C all form from their constituent parts in the absence of catalysis. The organizational formulae represented here are Ø → A, Ø → B, and Ø → C.

Because the Ø node is included in the SCC and all the nodes in an SCC are reachable from one another, there is no sink state in a graph of such a system. The state of the system will cycle within the SCC forever. The expected lifetime for every node in the graph is therefore infinity, and the *relative* lifetime for almost every node is, likewise, infinite (again, the exception is the Ø node).

Partial Spontaneity

We could also explore a system in which there are some patterns that are spontaneously organizing and some that are not. For instance, in A-B-C space, we might have Ø → A and Ø → B, but nothing that causes C.

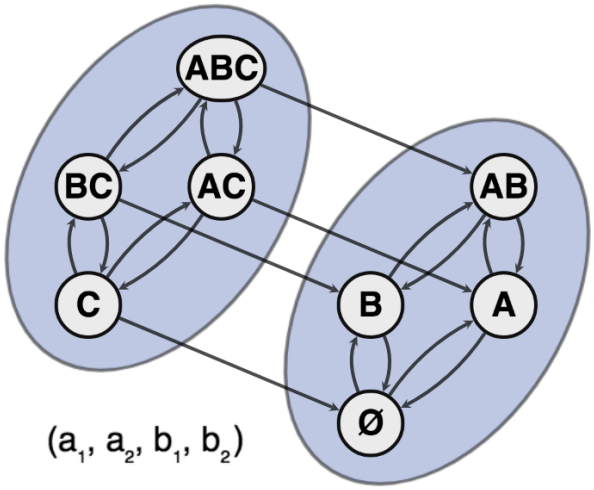


Figure 11.4: An organizational graph representing the spontaneous formation of A and B (from environmentally available foodstuffs, a_1, a_2, b_1, and b_2) and where there are no processes contributing to the production of C.

In this case, there will be multiple SCCs, one of which will represent the space of all the spontaneously organizing patterns in the system in the *absence* of any non-spontaneously organizing pattern, and then some of which will represent the same set of those spontaneously organizing patterns in the *presence* of one or some of the other patterns. For instance, in Figure 11.4 there is one SCC representing all the possibilities of A and B without C, and one representing the same with C. If we look at the SCC of just the spontaneously organizing patterns without any of the other

patterns, the lifetimes of the nodes in the SCC are like those in any standard spontaneous system—both expected and relative lifetimes are infinite for every node (except for the undefined relative lifetime of the Ø node). But if we look at any of the other SCCs, we will find that the lifetimes of the nodes in them will be much like those of nonce—no greater than the reciprocal of one of the patterns within the set that is susceptible to braising.

These kinds of systems are really just spontaneous systems analyzed in the presence of *irrelevant dimensions*. Despite the patterns in the irrelevant dimensions being susceptible to braising, ultimately the overall system will settle into the lowest fully spontaneous SCC where the lifetimes are infinite for that spontaneous subset of patterns.

Other Mixed Spontaneous Systems

Other systems might include some spontaneously organizing patterns along with some catalytic activity. It appears that these cases are generally similar to cases we have already seen. For instance, we might imagine a system in which A and B are spontaneously organizing, and B catalyzes C. In this case, the system turns out to have a single SCC encompassing all the nodes, and thus it behaves just like a standard spontaneous system, with infinite expected and relative lifetimes for all combinations of A, B, and C.

We might also imagine a system in which A is spontaneously organizing and A catalyzes B, but C is neither spontaneous nor catalyzed. This case is a hybrid of the case we just imagined and of the case of partial spontaneity that we looked at in Figure 11.4. The overall graph—and thus the behavior of the system—is isomorphic with that of the system in Figure 11.4.

In general, the behavior of spontaneous systems is that, in terms of those patterns that are neither spontaneously forming nor catalyzed by spontaneously forming patterns, braising will

eventually dominate . . . until what is left is just those patterns that either are spontaneously forming or are catalyzed by spontaneously forming patterns, at which point the system will retain its spontaneously organizing identity indefinitely. To be sure, this doesn't at all guarantee that any particular *physical* structure will last indefinitely—the physical state of the system might change over and over again, within the SCC. What is guaranteed is that, given an environment that contains certain assumed features, the total (actual plus potential) *organizational content* of the identity in that SCC will last as long as those environmental assumptions still hold. The blueprints for the patterns produced in the system live entirely in the assumed foodstuffs of the system and so, in the absence of those products, the blueprints remain, giving the products the chance to be produced again.

C. Teleological Systems

You see, proteins, as I probably needn't tell you, are immensely complicated groupings of amino acids and certain other specialized compounds, arranged in intricate three-dimensional patterns that are as unstable as sunbeams on a cloudy day. It is this instability that is life, since it is forever changing its position in an effort to maintain its identity—in the manner of a long rod balanced on an acrobat's nose.

—Isaac Asimov (1950)

Asimov was prescient. His recognition that life consists of a series of constantly changing states of a milieu of biochemical components, in order to maintain a living thing's identity is Kantian in nature and prefigures Maturana and Varela's work by about twenty years. It also looks just like our third category: the teleological systems. And so, last but not least, we can look at those types of identities that have some Kantian circularity in their catalytic capacities. The various autocatalytic sets—the little "miracles of self-reference" that we examined near the end of Chapter IX—all redundantly contain their own potential organization, and it is this organizational structure of mutually causal relationships that allows these identities to last longer than ontological nonce, but not as long as spontaneously organizing systems. We'll look at the details of how to quantify this in a few moments but, in short, the major SCCs in these systems always have finite expected lifetimes that are greater than baseline, and finite relative lifetimes that are greater than 1.

Standard Teleological Systems

A standard teleological system in A-B-C space is the version of Kauffman's autocatalytic sets that we analyzed in Figure 9.27, in which A catalyzes the formation of B, B catalyzes the formation of C, and C in turn catalyzes the formation of A. The graph of that system resulted in two SCCs—one containing just the Ø node, and the other containing the remainder of the nodes in the system. We can have another look at the same graph now in Figure 11.5.

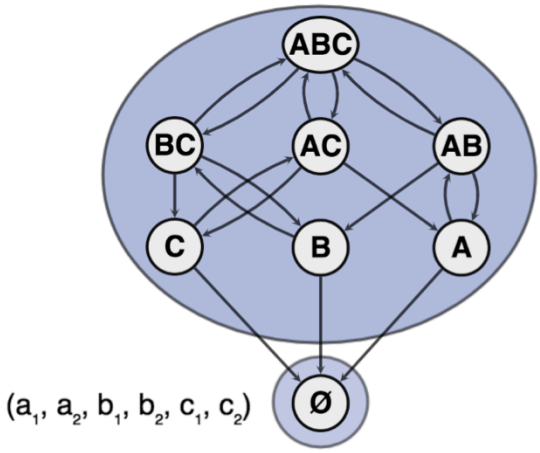


Figure 11.5: The autocatalytic set in which each of A, B, and C is able to catalyze the production of one of the other members.

In this system, the expected lifetime for the Ø-node SCC is infinite (and its relative lifetime is undefined), reflecting the fact that when all three of the catalysts disappear entirely, they are irrecoverable. However, for the larger SCC in this sort of graph, the expected lifetime of every node is higher, sometimes significantly higher, than the baseline lifetime. And for all of those nodes, the relative lifetime turns out to be finite yet greater than 1, revealing the fact that, in contrast with

nonce and spontaneous systems, these kinds of organizational identities are able to persist longer than a set of patterns would otherwise be expected to, although not forever.

Co-Catalyzed Teleological Systems

In the above autocatalytic set, we have a case in which A alone contains the organizational potential for both B and C, and in which B contains the potential for A and C, and C contains the potential for A and B. If the state of the system falls down to the node that contains just A, for instance, it is still within the identity, and so the other parts of the system can still be rebuilt. We should not, however, draw the conclusion from this example (and from that of multiquines) that any one part of a teleological system generally contains the full organizational potential for the rest of the system. That turns out to almost never be the case in real biological systems, and we can easily see that it is does not need to be the case in our graphs either.

As an example of a teleological system in which no one pattern ever contains the full organizational potential for the entire system, we can look at an autocatalytic set that requires co-catalysis (even if it is not chemically realistic). For instance, we can define a system in which A and B together are required to catalyze the reconstruction of any lost C, and B and C together are providers of the information for A, and so on. In such a case, nodes that contain single patterns (nodes A or B or C) are evicted from the uppermost SCC.

Figure 11.6 is a graph of such an example. The causal rules of the system are AB → ABC, BC → ABC, and AC → ABC, meaning that any two of the structures together are able to reproduce the third, but no single pattern contains all the organizational potential for the other two.

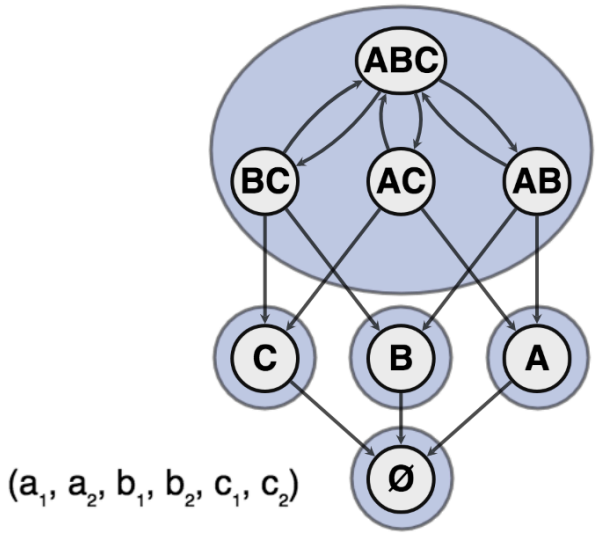


Figure 11.6: An organizational graph of a system wherein any two patterns are able to co-catalyze the production of a third.

In this system, as we might imagine, the expected lifetime of the Ø node is infinite, and the expected lifetimes of the other three singleton SCCs are the reciprocal of their respective decay rates (*i.e.* equal to baseline), while their relative lifetimes are all equal to 1. However, as with our previous example of a teleological system, all the nodes in the upper SCC here have expected lifetimes greater than their baseline and thus, also, relative lifetimes greater than 1. Again, the system will tend to cycle within the SCC, thereby maintaining the organizational potential for the set of patterns represented therein for longer than if the patterns did not have their mutually reinforcing catalytic capacities, but not necessarily forever.

Asymmetrical Teleological Systems

Lastly, we can look at a slightly differently-shaped system that is nonetheless autocausal. If we have a set of causal rules such as A → AB, A → AC, and BC → ABC, then the potential for B

and the potential for C both exist independently in A, and at the same time the potential for A exists in B and C jointly. The graph produced by this system is not symmetric, but it is nonetheless teleological.

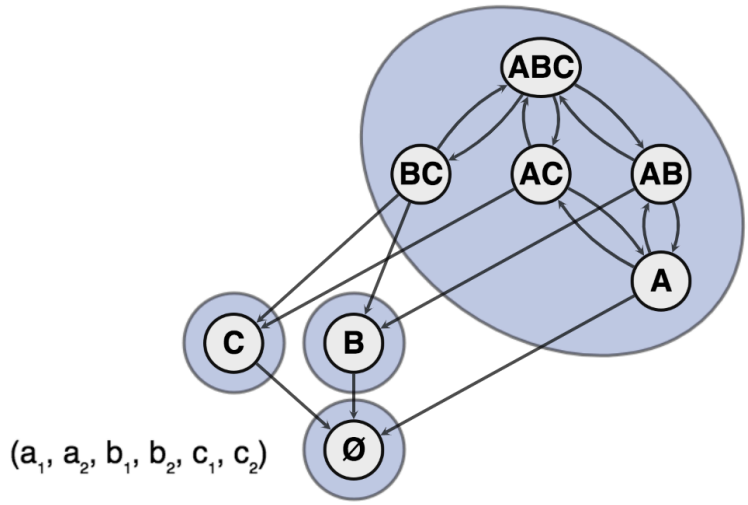


Figure 11.7: An organizational graph of a system wherein A catalyzes the production of each of B and C, while B and C jointly catalyze the production of A. The system is teleological.

The system in Figure 11.7 has four SCCs. The largest of them reflects the full set of catalytic relationships that exist in the system, and every node in it has an expected lifetime that is greater than baseline, and a relative lifetime that is greater than 1. The Ø node and the other singleton nodes all behave just as we would expect from our earlier analyses. This is a third kind of teleological system, in which the large SCC represents an identity that maintains its organizational potential for longer than would be the case if the relevant patterns had not had their mutually reinforcing catalytic capacities.

Helping Oneself: The Magical Transition

In Chapter IX we developed a theory of organizational identicality that provides a simple notion of a self. Now, with our lifetime calculations, we are able to see how certain identities—those that have internal organizational relationships that allow their parts to be both cause and effect of one another—are able to help themselves continue to exist longer than they otherwise would. It is the things that the interlinked causalytic parts of these systems do that cause the systems as a whole to persist. Unlike spontaneous systems, in which the persistence of the identity is driven solely by the organizational potential in the foodstuffs of the environment, in these teleological systems the persistence of the identity depends upon the causal relationships amongst those products of the system that also play a role in producing the system. Each pattern that is produced contains some of the redundant potential organization for producing some of the others. And so the system as a whole engages in a continuous series of acts, the upshot of which is the creation, maintenance, and reproduction of an organizational self.

Teleological systems are those systems that are neither guaranteed to exist nor guaranteed not to, but whose fate lies in their own hands. They are neither bound to last forever nor to assuredly fall apart (within a particular environment). Instead, they are little eddies in the universal flow of material organization. They are patterns in time. And they are able to benefit through the activities—now rightfully called *actions*—of their internal constituent parts, thereby contributing to their own good, meaning they are not just persistors, but also in fact agents, actors, subjects.

D. Nature's Blueprints

Concepts like those of organization, wholeness, directiveness, teleology, control, self-regulation, differentiation and the like are alien to conventional physics. However, they pop up everywhere in the biological, behavioural and social sciences, and are, in fact, indispensable for dealing with living organisms or social groups. Thus, a basic problem posed to modern science is a general theory of organization.

—Ludwig von Bertalanffy (1956)

We may not yet have explored all the possible kinds of organizational identity, but the categories we've come across so far provide an important foundation to any further exploration. We started with four different kinds of ontological nonce, all of which had different types of potential causal interactions among the patterns in their graphs, but which also shared a property: the full information necessary to build or rebuild those identities doesn't exist within the system (*i.e.*, within the identity together with its environment). Then we reviewed spontaneously organizing systems, where we also found some variation, along with another shared property. Specifically, we found that the blueprints necessary to build or rebuild these identities always exist entirely in the available environmental foodstuffs. Lastly, we looked at a few variations of teleological organization. In each of these variations, we found that the information necessary to build or rebuild these identities always exists in the system, but not just in the environmentally available foodstuffs—part of the information is located in some of the products of the system that turn out to be involved, directly or indirectly, in their own production. In these cases, the identity is partially responsible for containing its own blueprints.

It is my contention that these three major categories provide us with a full account of nature's blueprints. As a system of classification, the three categories show us what kinds of things can come to exist in a universe made of causally interacting patterns, and they also show us where the information that guides the production of each type of organization lives.

Insensitivity to Transition Probabilities

The reason I have been able to make generalizing claims about the resulting lifetimes of the nodes in SCCs without discussing the transition probabilities that generate the graphs that contain those SCCs is that those results are generally insensitive to changes in the transition probabilities. Formal proof of this insensitivity will have to wait until such time as I can put more work into this project. Suffice it to say that for this occasion, it can be shown that no matter what the likelihoods of the transitions are, the categorical lifetimes that emerge with our three organizational types go unchanged.

I think this is easy to see in the case of spontaneous organization. When the graph is a single SCC, one recognizes that the expected lifetime of that SCC is always infinite, owing to the fact that the state of the system can never leave the SCC. Changing the transition probabilities will affect which states of that SCC the system tends to spend most of its time in, but it will not change the facts that every state of the system contains the same organizational potential and thus that the expected lifetime of that organizational potential is infinite. In the case of ontological nonce—at least in the case of standard nonce—the expected lifetimes are equal to baseline lifetimes, and so they vary with transition probabilities, but still those expected lifetimes are always equal to the baseline (which also varies with the down-bound transition probabilities). The teleological case is more complicated, but the same sort of insensitivity holds. One can arbitrarily lower the probability

of up-bound connections, making them far lower than the down-bound connections of braising, and yet the expected lifetimes for the nodes in the autocausal set will always remain at least a tiny bit greater than the baseline lifetimes.

This insensitivity to the ratio between the transition probabilities of up-bound and down-bound connections shows that the feature that matters most to our categorical expected lifetimes is the organizational relationships within the system.

Tabulating the Cases

To make it easy to compare the behaviors of our three categories and their lifetimes, I here will tabulate the variations that we've explored so far.

Type	Expected Lifetime	Relative Lifetime
Standard Nonce	= Baseline	=1
Catalytic Nonce	= Baseline*	=1+
Flat Cyclical Nonce	= Baseline	=1
Causalytic Poisoning	≤ Baseline	≤1
Standard Spontaneous	Infinity	Undefined
Partial Spontaneous	Infinity	Undefined
Catalytic Spontaneous	Infinity	Undefined
Standard Teleological	> Baseline	>1
Co-Catalyzed Teleological	> Baseline	>1
Asymm. Teleological	> Baseline	>1

Figure 11.8. The major categories of material organization, including a few subdivisions within each type, and their expected and relative lifetimes. For detailed explanations of each of these cases, see the text in the earlier parts of the chapter. The unusual case of catalytic nonce is examined further below. The asterisk indicates the fact that the expected lifetimes of the nodes in these SCCs are equal to the baseline lifetime *of the underlying catalyst*, rather than their own baseline lifetimes. The plus-sign refers to the fact that the relative lifetime is 1 for *one of the nodes* in these SCCs, but it is greater than 1

for the other nodes (this contrasts with teleological organizations in which the relative lifetime is greater than 1 for *all of the nodes* in the SCC).

For most types of nonce, the expected lifetime of each of the nodes in any SCC in the system is less than or equal to the baseline lifetime for that node as a singleton; the exception is the case of catalytic nonce, where the expected lifetimes of all the nodes in the SCC tend to be equal to the expected lifetime of the catalyst that produces the other patterns in the SCC. The relative lifetime is then less than or equal to 1 for most nodes in most nonce SCCs, with the exception again being catalytic nonce, in which some patterns have their relative lifetimes mildly extended, thanks to a benefactor catalyst that produces those patterns but is itself vulnerable to braising.

We should take particular notice of this case. The "1+" listed in the table in Figure 11.8 refers to the fact that, in a purely catalytic (but not autocatalytic) identity, the relative lifetimes of the nodes vary; the relative lifetime of the underlying catalyst alone is equal to 1, but that of the other nodes will be higher. This is because catalysis extends the expected lifetimes of nodes with catalytic products, making them equal to the expected lifetime of the underlying catalyst.

Catalytic nonce thus turns out to be interesting, because those expected lifetimes are extended but, ultimately, they can never exceed the expected lifetime of the underlying catalyst. The set of patterns created by or rooted in the catalyst cannot be expected to exist any longer than the catalyst itself. This limit means that the identity is neither spontaneously organizing nor teleologically organized, but has a fixed, finite amount of time that it can be expected to exist.

Still, for the catalyzed product, this finite amount of time gives it a little longer than it would have had on its own. And this catalytic step upwards forms the building block from which teleological, autocatalytic sets and cycles are made. If a catalyst can hold its products' heads above water just a little bit, and if we can turn those products around and let *them* in turn act as catalysts,

holding the original catalyst's head above water just a little bit also, then the (now teleological) system can stay afloat for quite a bit longer.

If we return to the table, we can see that the remaining cases are all simpler to examine: In any type of spontaneously forming pattern, expected lifetimes are always infinite, and relative lifetimes are always undefined (the divisor of the quotient—the baseline lifetime—is infinite; again, see footnote number 343 where this issue is first addressed). And in the case of teleological systems, lifetimes are neither infinite nor limited by the braising rate; instead, the expected lifetimes for all nodes in a teleological SCC are always greater than baseline, and thus the relative lifetimes are always greater than 1.

The Tripartite Ontology

Presumably there are other as-yet-unexplored types of graphs containing organizational identities that may have as-yet-unimagined characteristics or behaviors that may prove to be interesting in various ways. However, there is a good reason that, for the time being, we can leave those stones unturned. Mathematically speaking, the temporal regimes of the categories we have found already completely fill out the spectrum of persistence behaviors. The lifetime of an identity is either (i) between zero and the baseline, or (ii) finite, yet above the baseline, or (iii) infinite. Any new type of organizational identity that may be found will inevitably turn out to be a subcategory of one of these three types. Such an identity will have organizational content that is either guaranteed to exist (it will be a new subtype of spontaneous organization) or guaranteed not to (it will be a new subtype of ontological nonce), or else its existential fate will be in its own hands (it will be a new subtype of teleological organization).

The byproduct of having developed our theories of identity and value, then, is that not only do we come to understand how a thing might be able to help itself, but we also come to understand the source of nature's blueprints.

E. A General Theory of the Nature of Living Systems

Insights gained from philosophical investigations into language and logic strongly suggest that the seemingly interminable nature of the controversy over life's definition is inescapable as long as we lack a general theory of the nature of living systems and their emergence from the physical world.

—Carol Cleland and Christopher Chyba (2002)

Psychologists have found that what is most salient to humans about death is that it is a universal, irreversible, non-functionality. That is to say: Everything that is alive eventually dies; and what it means to die is to cease to function; and this inevitable inability to continue functioning is permanent. Beyond a certain threshold, there is just no coming back (see, *e.g.*, Carter 2016). In contradistinction, there is what we call life. And so life can be thought of as a rare and temporary functionality. That is to say: only certain things of a certain sort may be alive. And what it means to be alive is to be able to have one's parts continue to function. And that ability to function will only ever last a finite period before coming to an end. As it turns out, this is the same characterization that emerges from our theory of informational identicality and expected organizational lifetimes (except that the theory also specifies to what end the parts of the living may function).[344]

The tentative theoretical definitions for the long-elusive concepts of *life*, *death*, and *health* that I'll offer now will probably not account for all the ways in which we think of those concepts, but I

[344] This intuitive characterization seems much like conceptual analysis. However, in this case the intuitions are not derived from a philosopher's introspection, but from a kind of psychological experimentation—now sometimes called "experimental philosophy"—in which an author derives the intuitions from the opinions of a population. The method is still very fallible, but the characterizations it produces can often be more balanced than just one philosopher's opinions.

think the offerings will be intuitive, and their rough outline might one day—if the rest of this theory pans out as I imagine—form a foundation for more scientific treatment of these notions.

I'd like to emphasize the fact that the theoretical claims regarding life that I am about to set forth emerge nearly effortlessly from what I've written so far—they come out of the few simple assumptions that my model is based on. We have assumed:

(i) that physical, organizational patterns play dynamically specific causal roles in one another's coming to be (whether by spontaneous or causalytic synthesis or decomposition);

(ii) that there is probabilistic decay in the material organization of patterns, and that that decay tends to accumulate;

(iii) and that identity consists not of *physically* but of *organizationally* identical states of the world (a theoretical definition supported by a distinction between *actual* and *potential* organization).

Using those ingredients, we have a small system for reasoning about the potential changes in organizational structure, from which a simple mathematical property—the relative expected lifetime of nodes—naturally emerges.

Life, Time

Of course, the term "lifetime" was chosen because it refers metaphorically to how long the organizational information within an identity might last, whether or not the identity is of the teleological type that we might view as vitalistic. But it has a stronger, more literal meaning when we

look in particular at those teleological identities whose behavior, when they do exist, helps them to continue to exist. The theory of life suggested here is this: to be alive is to be an active physical instantiation of a teleological identity. It is to be a thing that, thanks to the machinations of its parts in helping to reconstruct one another, is able to buy itself more time.

The sharp boundary drawn between an SCC and the remainder of a graph is, in these analyses, the boundary between the existence and the non-existence of the set of patterns in the potential organization of that SCC. All the states within the SCC have the same potential organization, the same identity. I suggest that what it means to be alive or not is for the actual state of the system to be on one side of this boundary or the other—to be in the identity, or not to be in it. This distinction differentiates the living not only from the once-teleological-but-now-dead, but also from the two kinds of never-animate matter that we have been calling spontaneous organization and ontological nonce.

Departure from the identity represents permanent death because, as soon as the SCC is exited, there is no longer a source of information, either in the identity or in the environment, for reproducing that organizational information.

Health

This definition of what it means to be (or not to be) alive also immediately suggests a measure of *health*, primarily in terms of informational redundancy. The more redundant organizational information the current state of a teleological identity has, the higher its chances of resisting braising. The further the current state is from any transitions that lead out of the identity, the less likely that state is to get bumped over that boundary by a random perturbation.

If we recall the earlier image we had when analyzing replication in terms of the draft of the iceberg, we can see that, in principle, a highly replicated system might get to be very far from the risky boundary of that identity, near the leading edges of the graph. In such a case the identity is very healthy, and thus highly resistant to potential braising. On the other hand, when the current state of an unreplicated autopoietic system is very close to the boundary of the SCC, it will be highly susceptible to a disorganizing perturbation, and thus very unhealthy.

Physical Realism

In the current instantiation of our system of analysis, time steps are measured in discrete *ticks*, a fact that keeps the model at a level abstracted from the physics that we take to be real (itself measured in terms such as kelvins, amperes, kilograms, meters, and of course *seconds*).

I have neither the time nor the capacity to translate the abstract model into a physically realistic one. But I can perhaps point out one way that we might eventually approach that problem. Each particular pattern that plays some causalytic role in a model will, in the context of a particular distribution of braising energy (within a particular environment), have a number of seconds that it can be expected to persist through resilience. Those details have been condensed, in our current system of analysis, down to a set of transition probabilities between nodes. But in principle, it seems that one might be able to reconceive the transitions of our graphs in terms of the real statistical expectations given by the distributions of energy and the likelihoods of causal interactions and so on, and thereby come to a more realistic model. If one day someone is able to do so, then I imagine it may be a significant step towards understanding the concepts of life, death, and health as physical phenomena, measurable perhaps in terms of information (bits) and time (seconds). In the meantime, the abstract version we have been looking at will hopefully serve us well in revealing

some of the key relationships between the high-level subjective concepts and phenomena that we set out to understand here.

Chapter XII

Final Thoughts

Time is the most valuable thing that a [person] can spend.

—Theophrastus[345]

Time is limited, even for those of us who temporarily are able to buy more. I had intended most of each of the following sections to be chapters of their own. As things stand, however, the parts of this project that come before this point are not yet complete. And so, writing those chapters that lie ahead will have to wait . . . perhaps a long time. For now, I will only put down a series of impressions about ways to interpret the graph theoretical methodology we've developed, and about some directions for future work. I will summarize what I think the theory is able to do, and I will explore some of what it is not yet able to do.

At this point, we have an abstract theory of teleology and of its subjective accompaniments, but along with those pieces comes a theory of the kinds of organizational structures that can emerge from the causal and temporal fabrics of a material world, accounting for some grand ontological divisions by which we might classify the diversity in the larger-than-particle-scale patterns found in our universe. It is not necessarily the only quantitative theory that can describe the phenomena in these branches of inquiry, and it does not describe everything about those phenomena, but it is—by my lights—a rather intriguing theory for describing some of the central aspects of those phenomena. And in spite of the oversimplifying assumptions the theory makes and the abstract level at which its most solid pieces have been described, some of its key strengths are its simplicity, its (purported, but

[345] As cited in Diogenes Laertius, Chapter V, Life of Theophrastus, X (Yonge 1853).

as-yet-unverified) explanatory breadth, and its consistency with our modern materialist logic and knowledge of the physical world.

To some eyes, the simplicity of the view I've described may appear to be a gross oversimplification of some of the deepest philosophical and scientific concepts that make up our physical, biological, and agentive lives. But to my eyes the theoretical concision is encouraging. I am not sure that Occam's razor or the principle of parsimony should guide our theories; but in this case, it seems justifiable that the obvious conceptual connections tying together identity, normativity, ontology, teleology, and agency should be mirrored by a hypothesis that is equally direct in how its theoretical parts are related. Our graph-theoretic system for describing the distribution of organizational potential within sets of patterns appears to be capable of making all these connections—deliberately and not haphazardly—with only a single central formula allowing us to calculate the expected lifetimes of the nodes in the SCCs of organizational graphs: $(I - P)\ EL = 1$. And as we have seen, the logic that holds all these pieces together may be abstract, but it is not abstruse.

The theory claims that the kinds of material order that can exist in the context of braising are those whose blueprints are, somehow, *available*. And the two primary ways in which blueprints can reliably be made available are either (i) as a result of a serial chain of spontaneous organization, or (ii) as a combination of some spontaneous organization along with some reflexive (Kantian) organization that continually constructs and repairs the redundancy in its own blueprints. That latter kind of material order forms the basis by which an identity—construed of as a *persistent organizational potential*—may benefit from its own actions, thereby laying the groundwork for subjective, agentive, teleological phenomena.

This set of concepts accounts for identity and value and autocausal persistence and, thus, for the goal-directed nature of teleological systems, and, as we'll soon review, also for functioning. It

also simultaneously supports four centrally biological capacities—autopoiesis, replication, and, as we will briefly review, also ontogeny and symbiosis—each of which is tightly linked to the notions of identity and value and constitutes a major part of what makes biological activity interesting. Furthermore, it supports the most fundamental property of biological entities: the distinction between, on the one hand, life or vitality (and the graded continuum of health) and, on the other hand, death or inanimacy . . . that universal, irreversible, state of non-functionality (Carter 2012). What we have in this theory is really a form of *emergent* vitalism.

But there is more to life than this. The breadth of phenomena that we find in organisms and ecosystems is certainly vitalistic, but it is far, far more complex than the descriptions given by the simple graph-theoretical models we've looked at. For one thing, while the abstract space of organizational potential does map onto the *coexistence* of physical patterns, I have not yet been able to show how to map it onto specific *arrangements* of coexisting physical patterns, such as that which occurs in cells or multicellular organisms (this complex issue was approached in our discussion of the chemoton, but it was not thoroughly answered there). For another thing, there are no vitalistic systems out there in the biological world that correspond to graphs symbolized in a space of just three letters. And I have not yet been able to make it entirely clear how to scale up such small graphs in a way that might allow them to represent real systems, including organisms. For a third thing, biological replicators can be far more complex than symbolic replicators or simple autocatalytic chemical replicators. Our graph-theoretic method of analysis may be able to obviously account for those latter types of teleological system, but quite a bit of work remains in order to determine how the same method might account for the kind of replication that occurs, say, in sexual reproduction with chromosomal crossover. I cannot address all of these topics now, but I will try to say a few words about those that I think I can.

One of the topics I would like to address in the next few pages is how our simple graphs may grow and combine open-endedly, in ways that mirror some of the complexity of biological reality. We will only be able to scratch the surface of this topic, but in looking at it, I will try to remind us of the role both chance and selection play in diversifying and adapting persistors within their environments. I will also point out how I think the kinds of symbiotic relationships that characterize both ecologies and individual organisms may be construed in terms of various graph-arithmetic operations between identities.

I also feel an obligation to remark on how the theory of teleological systems then impacts analyses of the concept and the phenomenon of function. In the early part of the dissertation, we spent a lot of time examining function theories. Although I do not have the time to persuasively defend a new theory of function, I will make some initial suggestions as to how we might now conceive of functioning in terms of teleological systems, leaving the fuller analysis of those issues for future work. When that is done, I will also be in a position to assess the degree to which my work addresses the twenty questions raised in Chapter VI.

In addition to all of these matters, I would also like to make note of another piece of work that I am aware of that makes some similar interpretations of some of the key concepts that I have been working with. Let's do that first.

A. A Similar Interpretation

Combining robustness and plasticity provides a measure of viability as average expected survival time under ongoing perturbation, and allows us to measure how viability is affected as the configuration undergoes transitions.

—Eran Agmon, *et al.* (2016)

A group of researchers, led by my colleague Eran Agmon in his graduate work at Indiana University, and including one of my own advisors, Randall Beer, have come to some similarly graph-theoretic conclusions about identity, value and vitality, which I would like to describe here.

Agmon and his colleagues (Agmon, Gates, and Beer 2015; see also Agmon *et al.* 2014, 2016) developed a cellular automata-like model of an autopoietic protocell in which there exists a membrane enclosure that contains an autocatalytic chemical, in an environment of both water and a readily available species of "food" molecule. Their protocell is built to model the diffusion of these chemicals across the membrane, and the formation of the autocatalyst and the membrane material from the food.

The authors' (2015) goal was to characterize the viability, ontogeny, and adaptivity—in other words, the health and life and the development—of an individual by exhaustively mapping the possible changes to that individual's physical structure. Their approach began by starting their protocell in a particular configuration and allowing it to proceed through time in the presence of random perturbations, in order that they could observe the structural evolution of the model over a broad range of possible scenarios. They then mapped the various states of change onto a graph meant to represent all the possible lives of that kind of protocell. It is not yet clear how to draw a

mapping between the kind of graph that Agmon *et al.* developed and the graphs that I have been describing.[346] Nonetheless, what is clear is that both are graph-theoretic analyses of similar topics.

Agmon *et al.* interpreted their graphs in some way similar to my interpretation of my graphs, in that *individuality* (their term most related to identity) and *value* are claimed to be features of the graphs (see the epigraph above); however, owing to the differing natures of our graphs, we've ultimately come to different conclusions about the meanings of those terms.

In their (2015) paper, Agmon *et al.* describe an "ontogeny"—a lifetime of an individual—as a trajectory through their graph, beginning at their initial state and ending at what they consider the "death" state (a roughly uniform concentration of each of their four molecular species across the entire space of the model). Trajectories of this sort cross through multiple strongly connected components of their graph along the way. Agmon's interpretation seems to be that an *individual* is best thought of as being an autopoietic *organism* (personal communication). I was personally unable to come to a clear understanding of whether an autopoietic organism is one actual trajectory (one ontogeny) or whether it is the full set of possible trajectories in Agmon's graphs (which he calls an "ontogenic network"), but in either case, these entities cross more than one strongly connected component. And while it is unclear, as I said, how to map Agmon's graphs onto mine, I find an interpretation in which identity does not map to some kind of clearly *organizationally identical* unit to be incongruous with my own theoretical interpretations.[347] As for the topic of value, Agmon claims

[346] Firstly, the representations within their nodes and mine are very different. Secondly, their explorations were strictly in a [model] physical space, while mine are in some kind of more abstract organizational space. And, thirdly, their graph is of what they call "stable configurations"—condensing series of transient states into singular nodes that serve as a kind of natural unit. It would take some work to understand whether those single-node units or the strongly connected components in their graphs are the clearest analog to the strongly connected components in my graphs.

[347] It may be the case that the SCCs in Agmon's graph turn out to be equivalent to SCCs in my graphs in some way. In that case, the nodes of each such SCC of his graph are organizationally identical with one another. We could then view his SCCs as being teleological identities that attempt to persist, and also view the transitions between his SCCs as being wildly damaging perturbations that result in still-viable alternative identities. In that case, those SCCs are not the same thing, but there may be historical paths from one to another. This way of putting things is somewhat similar to Agmon *et al.*'s own interpretation (since they classify *robust* vs. *plastic* responses to perturbations); however, it makes different claims from theirs about what constitutes an individual or an identity.

that value accrues to an individual—an organism—not to a more abstract informational or organizational entity, as I believe to be the case with the kinds of identities I have described. Because of this, we tend to disagree about the status of replicators as bearing the property of identity and as serving as potential beneficiaries.

Generally, the model that Agmon *et al.* have built impresses me, as an attempt to characterize the possible lives of an individual. Their protocell does appear to be a model autopoietic system of some sort, and so I remain hopeful that perhaps with further work we can one day come to understand how to reconcile their notions with mine. At the moment, however, it is not yet clear how that might be done.

Despite our inability to define our terms in quite the same ways, I am intrigued by a certain idea apparent in the work of Agmon *et al.*, whereby the ontogeny of an organism potentially may span a series of strongly connected components. There are two things that I think are worth exploring here. The first thing I'd like to explore is what it means to move from one viable organizational identity to another. The other thing I'd like to explore is the question of how my graphs might represent ontogeny—for example, the transformation from a caterpillar to a butterfly—even within just one identity. Let's explore the two issues in that order.

In Agmon's work, the notion of moving from one strongly connected component to another is a fairly obvious feature of his graph. In mine, the direct analog—a transition between SCCs made within a single graph—would be a destructive departure from an organizational identity. But another idea analogous to the move from one SCC to another in an individual graph could be represented by a parametric transformation of an entire graph into a different one within which a similar (but not identical) SCC might exist. One viable identity may transform into another that is no longer the same; nonetheless, the second identity may also remain viable. This of course happens to real replicators all the time in the course of evolution, both by drift (McShea and

Brandon 2012) and by natural selection. Systems change, and their denizens change with them. When a bacterium acquires some non-destructive mutation in its DNA, its daughter cells may still be complex teleological systems, but their identity—and the SCCs in the graphs that might describe them—will not be the same as those of the mother cell before the mutation. The new SCC here is not a *transition* from one part of a graph to another, but instead a result of the *reshaping* of the graph itself.

One interesting new question that is presented by trying to find the analogies between my work and that of Agmon *et al.* is the question of how often this kind of shift in identity might happen over the course of *autopoietic* persistence (rather than *replicative* persistence). We do know that organisms come to be infected by, or otherwise engaged with, various parasites and commensals all the time, thus changing both the organism and their relationship to their environment (Dawkins 1982). And in some interesting cases those parasites may come to functionally replace or functionally alter parts of the original organism (recall the tongue-eating louse from earlier, as reported by Brusca and Gilligan 1983). We also know that organisms might survive well, as they are, but are sometimes able to do better, in different ways, through the use of artifacts. In this case, the new part—the artifact that has been taken up—may come to play an unexpected role in maintaining the organism against braising, thereby redefining the organism's identity by serving as an additional part with causal effects that add nodes and transitions to a graph. And so there are at least a few ways in which the definition of an identity may be altered over the course of a lifetime, as an individual loses old parts or takes on new parts that begin to work in new ways with respect to the remainder of the original identity. For the time being, these ideas will have to remain inconclusive, but I think we must recognize that there is a lot of space still to be explored in coming to understand the notions of identity and value as they manifest in real organisms.

The second issue to discuss now is that of how a caterpillar and a butterfly can be seen to be "the same thing". Again, our analyses here cannot be complete, but we can look at how this general issue is addressed already in our graphs. If you recall, even within a single SCC in our organizational graphs, there are still potentially vast physical changes that a system may go through within one identity. The case of a multiquine makes it clear how two or more quite distinct looking programs can exist at different moments in time but still contain identical organizational information. And we can make the example more attractive: it is certainly possible to write a many-versioned multiquine half of whose versions maintain a fairly consistent (identifiable) form for a long period (with, say a counter embedded within, and decremented as it is passed from version to version), after which there might occur a transition to the other half of the versions which take a very different (also identifiable) form for a while before returning to the beginning. The "phenotypic" appearance of such a set of programs could be reasonably likened to that of a caterpillar and a moth or a hen and an egg. In fact, part of the point in developing the system of graphing that we did was to account for the changes in a physically diverse set of states that are, nonetheless, organizationally identical.

B. Commerce

Under a system of perfectly free commerce, each [entity] naturally devotes its capital and labour to such employments as are most beneficial to each. This pursuit of individual advantage is admirably connected with the universal good of the whole.

—David Ricardo (1817)

Individuals are not stable things, they are fleeting. Chromosomes too are shuffled into oblivion, like hands of cards soon after they are dealt. But the cards themselves survive the shuffling. The cards are the genes. The genes are not destroyed by crossing-over, they merely change partners and march on. Of course they march on. That is their business.

—Richard Dawkins (1976)

Let's discuss now the ways in which graphs, especially those that prescribe teleological identities, might be combined.

Biologists have for a long time now understood and documented the way ecosystems are constructed from complex networks of economic relationships between organisms that are all engaged in trade or thievery or scavenging of one sort or another. The same thing occurs at other levels of analysis.

For instance, while it has long been apparent that some of the relationships that exist within organisms may be similar to those in ecosystems, only in recent decades have biologists begun to shift towards understanding larger organisms not as being *unitary* identities but instead as being *collectives* composed of a network of smaller organisms interacting economically. The most clear and

thorough testament to this point of view is the emerging study of the microbiome—the vastly diverse community of microorganisms that make their home in and around multicellular organisms, in various states of relation, without which the multicellular organism may not be able to survive. But other evidence—*e.g.*, endosymbionts such as mitochondria and chloroplasts, as well as various studies of obligate ectosymbionts—all points to the same thing: individuals are generally not as individual as we imagine them to be, and they are generally not self-sufficient; most may only be able to persist when engaged in a series of commercial relationships by which they may make their living.

Richard Dawkins (1976, 1982) has taken the argument to a third level, beyond just seeing the individual in terms of an ecology made up of various organisms, and I think it worth following him here. Dawkins' idea is that the *genes* within an individual are the replicators that matter, and that they form a complex economy of their own—each vying for its own replicative survival, but cooperating with those it can, in order to earn gains from trade (Ricardo 1817). In light of our Kantian notion of an identity and a replicator, we can update Dawkins' idea, by viewing a gene not as an identity or a replicator by itself, but as potentially playing a role in the context of a more complete replicating teleological identity within the cell. If we view things in this way, then we might consider the genome—the collection of various genes that co-inhabit a "single" organism—to comprise a system of commerce in which various identities (if the relationships are facultative) or pseudo-identities (if the relationships are obligate) each bring something to the table in exchange for something else . . . and hopefully—for the sake of the entire community—coming to some economic equilibrium that helps the majority persist together, rather than undermining one another through excessive competition.

Graph-theoretically, we can see that, if A, B, and C are mutually catalytic or causalytic, and if A, B, and D are mutually catalytic or causalytic, then the full set of A, B, C, and D also form a teleological system, but the two identities (A–B–C and A–B–D) live side by side, partially

overlapping, sharing parts, and also, in some way, serving as an individual. Now, if we imagine A and B to represent the non-genetic machinery of cells and C and D to represent different genes that co-inhabit those cells, we can really understand Dawkins' gene's-eye perspective in terms of our teleological identities.

We can also look briefly at how a parasitic pattern could take advantage of a teleological identity. This could be just a bit of selfish-DNA (see, *e.g.*, Burt and Trivers 2006) or it might be a virus or a prion. In any case, let's imagine E happens to be catalyzed by A and B together, but it doesn't serve in any way to help create either A or B. If so, E can hitchhike on the backs of A, B, and C as long as C does the work to create the A and B that E also needs. If one reviews the graph of such a system (with the following rules: A + B → A + B + C, A + B → A + B + E, C → A + C, C → B + C) it turns out to be a teleological system wherein A, B, C, and E are all able to persist, despite E not providing anything for the others.

The upshot of all this discussion of the way commercial relationships operate at various levels, from the genome up to the ecosystem (and the actual economy), is that we can view our organizational graphs in two different ways. One of those ways is to make assumptions about which products or byproducts produced by other identities may be "freely available" in the environment and then, based on those assumptions, graph just the identity in which we are interested. The other way is that we can include two (or more) identities in the same graph and specify their coupling in terms of the ways the organizational contents of each affect the other. If we take this latter approach, we should be able to see which kinds of relationships might produce joint identities and which might not.

Although I am not prepared to spend the time drawing out in full detail how our abstract graphs might support analogs of each of the biological capacities of mutualism, commensalism, and parasitism, it looks as if such analyses will be rather straightforward. We have just done part of that

analysis (not to mention that we also began to look at some of the variety in couplings between potential identities when we graphed Eigen's hypercycles in the latter part of Chapter IX). I am hopeful that a deeper exploration of the various possible couplings between identities, in terms of the arithmetic operations that may combine or extend graphs, might be able to explain the spectrum of symbiotic relationships from mutual to parasitic. And that analysis may also be able to explain how those relationships may come to be *obligate* or *facultative*, *permanent* or *temporary*, and *cooperative* or *competitive*, not to mention the ways that multiple identities may even overlap and share physical parts—each using the same machinery for their own needs in the same or different ways. Analyzing graphs of the relationships between identities in this manner may help us use those graphs to account for many of the complex ways that the underlying notions of vitality, identity, and value come to be expressed throughout numerous levels of the biological and economic worlds.

C. Chance and Other Conditions

Chance alone is at the source of every innovation, of all creation in the biosphere. Pure chance, only chance, absolute but blind liberty is at the root of the prodigious edifice that is evolution... It today is the sole conceivable hypothesis, the only one that squares with observed and tested fact.

—Jacques Monod (1971)

Another thing worth exploring in future work will be the role that chance plays in the organization of systems. A system, as I've described them, is a temporary definition. It is a summary of the larger organizational patterns that might emerge from the causal relationships between various possible sets of organizational parts. But as an environment shifts and changes, system definitions change with them, and so, too, do the identities in those systems. It will be no small feat to try to understand what kinds of useful generalizations one might be able to make from this.

Not only this, but a system definition, specified only in terms of organizational content, is not enough to predict the behavior of material in a system. In order to get micelles to spontaneously organize, for instance, we require not just the right parts (the surfactants along with the solvent that attracts one of their ends), but also the right concentrations of those parts in the solution, and the right temperature and pressure for the system and so on (van Doren 2007). Chemical systems in general require energetic conditions that go beyond the potential in the reactants in order to proceed. For example, temperature is a factor for the catalytic processes that contribute to teleological systems too. Most real biological enzymes work only in a quite narrow range of temperatures. In humans that range is around 37°C, plus or minus a few degrees. The

physiological phenomenon of fever is so dangerous to us because, when our body temperature lies outside the effective range for the many enzymatic processes in our cells, the functional behaviors of many of those parts just don't work, and our autopoietic systems can no longer operate. So, in a system where the organizational relationships for a particular teleological identity are right, other conditions may still not be right and the teleological identity will never be actualized.

Perhaps all that this means is that the nodes in our systems need to be made more complex—they need to represent more about the causal interactions between parts—in order that we can say more about what really accounts for the transitions between them. I would be pleased to see if someone can discover a way to do just that. But it is not an easy job. Of course, were we to take up the task, I would find it crucial to respect the notion of redundancy that underlies persistence against braising, in order to retain the underpinnings for identity and value in our method of analysis.

Another issue that this discussion of further conditions brings to mind is that of how a teleological identity might come to exist in the first place. If these little eddies in the causal fabric of material order work so hard in order to buy themselves more time, yet inevitably may succumb to braising—if life is inherently so risky—then what causes them to form in the first place? From where do they come?

I think the largest part of the answer here is that we need to look beyond the theory of teleology, to the theory of evolution by descent with modification. As I said in Chapter V: natural selection *does not* make things teleological (that status, I claim, exists thanks to the Kantian structure of those things), but natural selection *does* make teleological things. It has the power to take existing teleological systems (replicators, especially) and modify them (by *chance*) providing the opportunity for some of those modifications to also exist as viable teleological systems (while others disintegrate into ontological nonce).

D. Organization, More Broadly

Who could foresee the organization of living beings, if the cellulose, which is right, should become left, if the left albumen of the blood should become right? There are here mysteries which prepare immense labours for the future, and from this hour invite the most serious meditations in science.

—Louis Pasteur (1860)

Localization of function is the law of all organization whatever: separateness of duty is universally accompanied with separateness of structure: and it would be marvelous were an exception to exist in the cerebral hemispheres.

—Herbert Spencer (1855)

The quotes from Pasteur and Spencer, about molecular chirality and the functional correlates of mental processes in the brain, remind us that organization is not always just the presence of organizational parts with some likelihoods of interaction. Organization is often a more complicated arrangement of spatial, structural relationships; it doesn't depend only on *what* parts there are, but also on *where* they are and just *how* they might interact with one another. While the theory we've analyzed discusses "organization" at one level, there seems to be another level at which the organization of a system needs still to be addressed.

This is the problem that vexes me most about the theory put together in this dissertation: It is clear that an autocatalytic set and other such diffusively interacting chemical systems may be interpreted straightforwardly by the kinds of organizational graphs we have looked at, but it is not

clear how more complex organization may fit the same mold, in order to justify the leap from the *abstract* model (that is strongly analogous to biology) to a *true* account of biological behavior.[348]

I dearly hope this problem is not insoluble, but unsuccessfully pondering it in search of a solution has troubled me greatly in the years since first developing the graphical approach to organizational redundancy.

The two paths that I think most likely to bear fruit in resolving this concern are either a richer understanding of the complexity that can be modeled by graphs such as those at which we have been looking, or a richer interpretation of the ways the contents of nodes in our graphs are able to influence one another. In the first case, I imagine that perhaps a sufficiently complex graph of the possible states of a system can model the parts of a system as well as their arrangements. I have, however, struggled to find a way to see our graphs as operating in such a manner. In the second case, I imagine a new interpretation of our nodes, in terms of both the parts that make up a structure and their spatial relationships, might allow us to specify the transitions between those nodes more carefully. This seems more tractable and comprehensible to me; however, I have still struggled to find a way to simply model all the details within a node, in order to realize an improved graphical method. In either case, it is clear that *some* additional theory of organization is required in order to make full sense of the ways the parts of systems interact and in order to bring this theory closer to a complete account of biological reality.

In spite of this vexing problem, I would like to point out the reason it is worth holding out hope: The theory, if it can be drawn out into something more than its current abstract version, presents us with an internally consistent and logically satisfying solution to a series—or network—of even more vexing philosophical problems. For that reason, I think that whatever form an eventual

[348] By the way, it should be noted, that this problem is not a problem for Agmon *et al.*, as described above; their work is an actual functional model of autopoiesis with all the necessary organization built right into the model.

complete theory of these topics takes will need to preserve the roots that I've developed here, regarding the notions of redundancy and organizational identicality.

E. Function Theories

In every context where functional talk is appropriate, one has also to do with the goals of some goal-directed system.

—Christopher Boorse (1976)

I suggested in Chapter V that, were each of the six main theories of function to make some adjustments and concessions, they might all come to converge on a unified story that somewhat resembles each one of those theories and that, more importantly, accounts for most of the observations that drove theorists to each of them. The concessions to be made in some cases turn out to be large but, in the context of what we have seen in the second part of the dissertation, and in the context of what is gained by such concessions, I think the amendments to the theories are justifiable. I will make some proposals now, beginning with the most obvious cases.

Functioning that is a causal contribution to the persistence of a teleological identity is almost literally what Christopher Boorse had in mind with his Goal Contribution analysis. The major difference being that Boorse intended his theory to be one of both strong and weak function statements—both ("has a"-style) proper functions and ("serves a"-style) functioning—while the current theory abjures the former and only directly addresses the latter. *If* Boorse recognizes—as he seems almost to do—that strong function statements are a product of the illusion of function constancy, *and if* he relinquishes the cybernetic theory of goals (which I think he is prepared to do, given a suitable alternative) *and if* he instead accepts my new theory of what makes for goal-directedness (which I cannot say if he would be prepared to do; but he might), *then* his theory of functions would seem to account for all functioning (as it already did), yet it would be given in terms that are no longer so vaguely defined. There is much work that would still have to be done to show

how the numerous examples and counterexamples of a conceptual analysis might hold up to this version of Boorse's theory, but I remain optimistic that his previous work could be amended to show how that might be done in order to account primarily for a central sense of the concept of functioning in terms of direct or indirect causal contributions to the persistence of a teleological identity, and then for a series of more peripheral senses of functioning that are metaphorical extensions of that central sense. One significant theoretical piece that would need to be worked out to make this work is the way in which psychological goals derive from biological teleology. The solution I propose is that the brain is a heuristic device meant, like any other organ, to help an organism achieve its biological goals. Psychological goals are attempts made in that direction. Some of those attempts succeed, some don't; and sometimes the system becomes hijacked and instead serves the goals of some other identity that gains control over it.

The same theory of functioning that I hope would satisfy Boorse is also closely aligned with Mark Bedau's Valuable Effects analysis. Simply put, to make a causal contribution to the persistence of a teleological identity is to have a valuable effect on that identity. Bedau would have to relinquish his commitment to the function–accident distinction, and he would have to put his distinction between grade-two and grade-three teleology (mirroring the distinction between biological and psychological goal-directedness) in different terms (here again I have in mind the idea, mentioned above, that psychological goal-directedness derives from the way minds *heuristically*, and fallibly, serve biological functions). But if Bedau did so, and if he accepted our new account of how value arises naturally from the persistence of organizational identities, then he may be able to retain the claim that value lies at the fundamental core of the notion of teleology (and functioning) and still account for most or all of the functioning we observe, while no longer being committed to a nonspecific notion of value. Again, I think quite a bit of work is required to make it clear how this would all occur, and it is not yet clear that Bedau himself would want to make those concessions, but I remain

optimistic that an account of functional items in terms of their valuable effects upon teleological systems may one day be possible.

Before moving on to the three popular theories of function, we can also review Mayr's Programmed Effects analysis. Although Mayr's analysis performed poorly with respect to counterexamples earlier, I felt there was an important kernel of truth in his, admittedly vague, notion of programming. We can try to fish out that kernel now. His tentative definition of programming as "coded or prearranged information that controls a process (or behavior) leading it toward a given end" (Mayr 1992) *could* be interpreted to mean that it is the organizational relationships between the parts within an item that make that item and its parts functional. And this interpretation *might* be seen as aligning fairly closely with the notion of teleological identities that we have been discussing. If the various states of an identity are thought of as containing different versions of the same "code" in the causal proclivities of their organizational potential (the same way the different versions of a multiquine do) then Mayr's idea of functional, teleonomic systems having programmed effects is a bit more coherent than I had previously analyzed. The organizational potential in the identity can be thought of as the prearranged information that controls the processes and behaviors of the system that continually guide it in the pursuit of persistence (the "given end"). This is still not a *very* clear interpretation of "programming", and I don't want to defend it strongly, but it does provide a way to generally understand Mayr's intuition.

We can review the three popular theories now. When we first looked at the Causal Role analysis, I used an accounting of examples to stress that the analysis seemed to reliably, but silently, require teleological systems as the ultimate context for the causal roles that justify function attributions (and that all the causal roles that we don't consider functional occur in the absence of teleological systems). Cummins and other CR theorists expressed their ideas in staunchly anti-teleological terms because of their conviction that the normativity of teleology is in direct

contradiction with their materialist beliefs. I agree with their materialism, but I think the theory of teleology and normativity that I've developed is in fact fully naturalistic, and so I would implore Cummins, Davies, and any other CR analysts to reconsider that there may in fact be a source for the "norms of nature"—a single concession that would allow their causal role theory to otherwise make perfect sense.

When we looked at the Selected Effects theory of functions, I was impressed with the fact that these theorists, in contrast to the CR theorists, *were* committed to accounting for the normative nature of functioning. I think that commitment was the core truth in the search that found the SE theorists arriving at the comparative, historical norms of natural selection. To those theorist's eyes, history appeared to be the only available norm by which to ground our observation that biologically functioning items ought to do whatever it is they do that we consider to be the function of that item. However, I didn't find those comparative norms satisfying; they seemed to me, for many reasons, to be inappropriately normative in accounting for functioning. Perhaps the foremost of those reasons was that historical norms force us to choose between believing in functions that are causally irrelevant and believing in some kind of historically granted residue such as a functino. In light of a theory of an emergent kind of *evaluative* normativity, however, I think those theorists who are committed to the normative nature of functioning might be able to back up and think about whether these norms can do the job better than historical norms and still make sense of not just why a thing may be functional, but also the question of "why is it there?" The answer I would give to that question is that functional items are there, *in the functional sense of the question*, because they are components of systems that work to produce themselves, regardless of whether selection has worked to alter them. (Many of them are also there in part because of selection—but that is a different sense of the question.) In the case of the SE analysis, I think the concession to be made to align with the theory of teleology would be larger than in most cases. One would have to generally

abandon many of the details of the SE theory, and hold on only to the driving factors for which that theory was established. That may be too big a leap to ask its proponents to make. But I believe if Millikan, for instance, were to take such a leap, she could still ground her biosemantics (her theory of meaning in biological systems) in the notions of functioning and its contribution to goal-directedness. While the details of making the theoretical adjustments clear would take a significant amount of work, I expect the payoff to be worthwhile.

Lastly, we can look at the Replication Dispositions analysis of function, which I suggested earlier would turn out to be very close to the theory of functioning in terms of teleological systems. The traditional replication dispositions analysis was given in two primary forms, one of which took functioning to be a causal contribution to the disposition to replicate and the other of which took it to be a contribution to the maintenance of an autopoietic system. The theory in terms of teleological systems agrees with both, because at an abstract level both replicators and autopoietic systems are forms of persistors that maintain their organizational redundancy. And so the update I would propose to the RD analysis would be simply to join its two branches, thereby framing functioning as a contribution to the disposition of an identity to persist. If earlier RD theorists were to concede that dispositions toward either replication or self-replication (autopoiesis) can ground functioning, and if they were to loosen their account and also admit indirect contributions—such as those made by artifacts, then I think they could work to eventually find their theory accounts for the majority of central cases of functioning, now including the counterexamples that their theory previously struggled with.

Ultimately, I think a lot of work still remains to be done to clearly understand the concept of functioning and the shape of the conceptual halo that characterizes it. But it seems clear to me that a theory of teleological systems that produces a natural source of normativity should deeply affect the assumptions of the prior six function theories and incentivize making some revisions to each.

F. Reaching the Goal?

Teleology is not the antithesis of causality, but subordinate to it. It is, of course, inadmissible to consider "final causes" as implying that an object or end is capable of having effect. No event that has not yet taken place can possibly act. But results are caused by the keeping of the end in view*, and it is in this way that the final becomes an efficient cause.* These final efficient causes are not in the slightest degree metaphysical*, for they derive from organic matter.*

—Hans Driesch (1914, emphasis added)

Earlier, I attempted to set some standards of achievement—in the form of my twenty questions—that I suggested should apply not only to the theoretical work here, but also to any future work in the same vein. The idea is simply that we can gain some initial confidence that a theory is on the right track if it demonstrates a moderate degree of explanatory completeness with regard to teleological phenomena, including both of the main streams of teleological observation—that of goal-directedness and that of function. If we find the fairly concise terms of our new theory to have broad explanatory power, then we should find ourselves at least tempted to unbox that theory further and see what else its pieces might be able to do for us and how we may be able to develop it further. As it turns out, we will have to leave the bulk of that unboxing to future work, but for now we can measure up what we've got so far by seeing how it performs in answering the twenty questions. I apologize in advance, because I find it best to address these questions out of the numerical order in which I had previously put them.

The first three questions inquired into the nature of teleology, identity, and value, and the fourth question asked how we may clearly draw a line between teleologically organized systems and

other systems that may have *some* similarly lifelike properties (such as the growth, healing, and metabolism that we find in spontaneously organized systems). I hope it is obvious at this point that the abstract theory I've given is a direct attempt to answer each of those questions. Again,

> (i) an identity is a set of states of a system that all contain the same organizational potential—that is, they can transform into one another freely;
>
> (ii) persistence is good for an identity and so things that have (subjective, relative) value are things that contribute to the persistence of an identity;
>
> (iii) teleology occurs whenever some identity is able to provide value to itself—whenever a system that contains redundant organizational potential is able to work towards increasing its own organizational redundancy and thereby work against the environment of braising that has the tendency to reduce organizational redundancy; and
>
> (iv) this can be put in contrast with spontaneously organizing systems, which are those that also sometimes grow, heal, or replicate, but do so purely through the machinations of environmentally present organizational contents, rather than the internal machinations of the very patterns that are growing, healing and replicating.

Systems that are teleologically organized are a subset of those dissipative systems that metabolize free energy to construct their patterns, and they are a subset of those systems that display perseverance and plasticity.

The fifth question asked what life itself is. And again, I think it straightforward to say that while our theory of teleological organization is in no way a detailed theory of cell-biological life, it does constitute a theory of vitalistic processes more broadly, and so, while much work remains to be

done, it can nonetheless be seen as abstractly encompassing not only cell-biological life but also other potential prebiotic or abiotic forms of vitalistic, teleological organization.

We can jump ahead now to the last seven questions (numbered fourteen through twenty), each of which asked whether a new theory of teleology could help us understand why some central aspect of an older function theory seemed to be so well correlated with items that function as to convince a theorist to propose a theory in terms of that aspect. The answers to the first six of these questions were addressed, in part, in the previous subsection, and there is no need to repeat those thoughts here. I don't think the answers I gave there are yet entirely satisfying, but I remain optimistic that, with more time and thought, the details may be worked out. The fact that Millikan (1984) saw *reproductively established families* as being the locus of the norms that cause functionality (question 20) is accounted for by the fact that goal-directed systems are, generally, reproducing items.

The discussion of those function theories above also provides some approximate answers to question seven ("what is a function?") and questions twelve ("can we account for accidental functioning?") and thirteen ("how can we reform our concept of functioning?"). The reformative aspect here is accounted for by the argument I gave in Part I that there are no proper functions (TDHF) and that we need to account only for function*ing*. Under that perspective, the question "what is a function?" is then replaced with "what is functioning?", which we can answer pretty much in the way Christopher Boorse has done. This makes it easy to account for accidental functioning too, because an item that happens to help a persistor persist for a bit longer than it otherwise would is, for that moment, serving a function.

There are now five questions remaining—numbers six, eight, nine, ten and eleven. I'll try to quickly answer the last three of those first, and then we can finish up with the other two, which are more interesting. The function statements found in the list on pp. 205-207 (question nine) are

generally accounted for by a theory of (verb-) functioning. Things that serve a function (FS10, FS11, F13, FS14, FS16, FS20) are directly accounted for; things that have a function (statements FS1 through FS9, FS15) only seem to have their functions because either they regularly serve functions or they are usually thought of only in terms of the functions that they may potentially serve; and the unusual function statements such as FS17 through FS19, are cases in which we have stretched the meaning of the term, either by conceiving of the functional item in a teleological context (FS18, FS19) or otherwise extending the conceptual halo of the term by analogy (FS17) (see Hofstadter and FARG 1995; Hofstadter and Sander 2013).

The many senses of "for" (question ten) can be unified under the view that ultimately being *for* something means being good for some teleological identity. In each case of being designed for, being used for, being meant for, and so on, we can trace the meaning of "for" back to the contribution of value to a teleological identity, in one manner or another.

Design (question eleven) can be thought of as any process that contributes to the construction of a thing that is able to function. The process is often—and probably best—performed by the hill-climbing procedure that Dennett (1995) calls "generate and test". But some things might become designed by just a single iteration of the process—one attempt at generation that produces a successful model. What ultimately matters is whether the item eventually comes to be functional. In some cases things might be poorly designed, and in other cases design-efforts may fail to produce anything functional, but when something is eventually determined to actually be able to function in providing some value to a goal-directed agent in some way, then the thing has been successfully designed. (Of course a designed item may never actually serve a function, but at least it is *able to*.)

We can look at our last two questions now. Question number six asked how we might address the problem of backwards causation. I take it that the seeming causal paradox that teleology

raises (whereby a thing in the future—a goal state or event—seems to be a partial cause of something in the present or past) is clearly resolved by the circular logic of teleology. Rather than avoid backwards causation, we can embrace it. As Driesch put it in the epigraph to this section, *results are caused by the keeping of the end in view*. Of course, we already recognize that this is true in our own human goals, wherein the process of striving toward a psychological goal is caused by our mental representation of an end state, rather than by the unrealized end state itself. But in biological goal-directed systems, results—future states of the system—can also be said to be caused by *keeping the end in view*, just as long as one understands that the potential organization for future states is built into the current state. There is a Q that's in P, and there is a P in the Q that's in P. Since every state of an identity is organizationally identical, every state of a teleological system *contains a representation of the system's intended future*. For an identity, there is a sense in which past, present, and future are all the same, and so the goal of persistence is always *tacitly* kept in view. The goal is represented (non-psychologically) by the identity itself, for as long as it persists.

A few lines up, I used the word "intended", in the teleological sense in which an intention is an aim towards a goal. But I would like the reader to also take note, here, of the relationship to the philosopher's notion of *intentionality*, which is synonymous with the terms "meaning" or "aboutness". When we talk about a goal being represented within the organizational potential of a system, that organizational potential comes to be quite literally *about* the future, and *about* the goal. The sense in which the potential encoded in an organizational identity is about the future is, I claim, the proto-intentional foundation upon which *anything* may come to be about anything. It is the foundation for the minimally cognitive behaviors of all simple living systems, and the eventual foundation for the ways in which the intentionality in minds comes to serve the goals of the complex communities of persistors that use those minds.

Our eighth, and now final, question had to do with whether teleology can provide for the autonomy of biology among the sciences. When introducing the notion of the autonomy of biology in Chapter IV, I already hinted at my answer to this question. I believe that there are three astonishing and fundamentally irreducible biological phenomena: first that there are living things, second that there is such a diversity of living things, and third that the diversity of living things are so well adapted to their environments. The second and third are both explained in some way by natural selection. If we divide natural selection along the lines that McShea and Brandon (2012) do, we can see that random, diffusive, genetic drift can begin to account for the diversity of living things, while selection can account for the rest and, at the same time selection can also account for the commonly found adaptedness of those diverse living things. I contend that a theory of goal-directedness—be it the theory I am offering now or some improved or alternative version of it—will be able to account for the first phenomenon, the existence of living things. A theory of teleology accounts for the characteristics that are central to life—namely striving, agency, projectivity, and vitality. At the same time, the particular theory of teleology that we have been looking at rules out older conceptions of vitality, such as those based in metabolism, reproduction and arbitrary boundaries such as cell walls, firstly by showing how to distinguish between teleological systems and spontaneous systems and, secondly, by abstracting the generic, cyclical, Kantian causality of teleological systems away from the specific metabolism–membrane autocausality that we find in cell-biological life.

If I am right in all this . . . and I understand it is not yet easy to measure whether I am, then those who find the functional nature of biological phenomena to be irreducible (e.g. Laubichler 1999) are close enough to right, and the biological reductionists are just wrong. I think that existence as a teleological system is the most basic property that accounts for the vitality we see in living things, and it is precisely what makes them alive. And since this property is emergent from a

certain kind of *organization* amongst causal relationships, it does not have its basis directly in the physical and chemical parts from which those teleological systems may be made. Teleology is truly an irreducible law of life, logically prior to both drift and selection, and thus a fair contender for the position of Biology's First Law.

So there are our twenty questions. On balance, I think I have given some kind of answer to each of them, but I don't think I have yet fully and satisfyingly addressed all of them. I have made strong progress towards some, and more impressionistic progress towards others. Although the results so far are inconclusive, I am encouraged to think it worth pursuing this line for two reasons. First, the majority of that inconclusiveness just calls out for more work; there are ways in which further progress might be made towards answering each of these questions within the current framework. And second, it *is* satisfying to imagine that such a diverse set of historically vexing questions might all be answered by a single, relatively uncomplicated, theoretical model. For all its incompletions, I contend that the modern theory of teleology, the main pieces of which have been around for fifty years but which have now been integrated into a fairly cohesive story, is able to bring us a step closer to a science of the subjective.

References

Achebe, C. (1958). *Things fall apart.* London: Heinemann.

Adams, F. (1979). A goal-state theory of function attributions. *Can. J. Phil.* 9, 493–518.

Agmon, E., Gates, A.J. & Beer, R.D. (2015) Ontogeny and adaptivity in a model protocell. *Proceedings of the European Conference on Artificial Life 2015*, pp. 216–223.

Agmon, E., Gates, A.J., Churavy, V., and Beer R.D. (2014) Quantifying robustness in a spatial model of metabolism-boundary co-construction. *ALIFE 2014: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, pp. 514–521.

Agmon, E., Gates, A.J., Churavy, V., and Beer R.D. (2016) Exploring the space of viable con gurations in a model of metabolism-boundary co-construction. *Artificial Life*, 22 (2): 153–171

Al-Ghazali, Abu Hamid (2001) کیمیای سعادت (*Kimiya-yi Sa'ādat*). Translated from the Persian by Muhammad Asim Bilal; Revised by Munir Ahmad Mughal. Lahore, Pakistan: Kazi Publications.

Allen, C. (2002). Real Traits, Real Functions? In Ariew, A., Cummins, R., and Perlman, M. (eds.) *Functions: New essays in the philosophy of psychology and biology.* Oxford: Oxford University Press.

Allen, C. and Bekoff, M. (1995a). Function, natural design, and animal behavior: Philosophical and ethological considerations. In N.S. Thompson (ed.) *Perspectives in Ethology, Volume 11: Behavioral Design*. NY: Plenum Press, pp. 1–47.

Allen, C. and Bekoff, M. (1995b). Biological Function, Adaptation, and Natural Design. *Philosophy of Science*, Vol. 62, No. 4, pp. 609–622.

Allen, C., Bekoff, M., and Lauder, G. (1998). *Nature's Purposes: Analyses of Function and Design in Biology*. Cambridge MA: MIT Press.

Amundson, R. (1996). Historical Development of the Concept of Adaptation. In Amundson, R. and Lauder, G. *Adaptation*. Academic Press.

Amundson R. and Lauder, G.V. (1994). Function without purpose: the uses of causal role function in evolutionary biology. *Biol Philos* 9(4):443–470.

Aquinas, Saint Thomas (1912) The "Summa theological" of St. Thomas Aquinas. London: Burns, Oates & Washburne, Ltd.

Ariew, A. (2002). Platonic and Aristotelian Roots of Teleological Arguments. In Ariew, A., Cummins, R., and Perlman, M. (eds.) *Functions: New essays in the philosophy of psychology and biology.* Oxford University Press: Oxford.

Ariew, A. (2007). Teleology. In M. Ruse and D. Hull (eds.) *Cambridge Companion to the Philosophy of Biology*, Cambridge: Cambridge University Press.

Aristotle, & Ogle, W. (1911). *De partibus animalium*, tr. Oxford: Clarendon Press.

Aristotle. (1924). Aristotle's *Metaphysics*. Oxford: Clarendon Press.

Aristotle. (1970). Aristotle's *Physics*. Books 1 & 2. Oxford: Clarendon Press.

Aristotle, & Peck, A. L. (1943). *Generation of animals*. London: William Heinemann.

Aronson, J. L. (1971). On the grammar of "cause." *Synthese*, 22, 414–30.

Ashby, W. R. (1947). Principles of the Self-Organizing Dynamic System. *The Journal of General Psychology* 37 (2): 125–8.

Asimov, I., (1950). *Pebble in the Sky*. Garden City, Doubleday.

Atkins, P.W. (1984). *The Second Law*. New York: Freeman & Co.

Atkins, P.W. and dePaula J. (2006). *Physical Chemistry for the Life Sciences*. New York: W.H. Freeman and Company.

Augustine of Hippo, & Dyson, R. W. (1998). *The city of God against the pagans*. Cambridge: Cambridge University Press.

Autenrieth, J.H.F. von (1836). *Ansichten über Natur und Seelenleben*. H.F. Autenrieth (Hrsg.) Stuttgart, Augsburg: Cotta.

Averroës (1954). *Tahafut Al-Tahafut (The Incoherence of the Incoherence)*. Translated from the Arabic with introduction and notes by Simon van den Bergh. London: Luzac.

Avicenna, Ibn Sina; Laleh Bakhtiar (1025). *Canon of Medicine*. New York, NY: AMS Press, Inc.

Ayala, F.J. (1968). Biology as an Autonomous Science. *American Scientist*. Vol. 56, No. 3 (Autumn 1968), pp. 207–221.

Ayala, F.J. (1970). Teleological Explanations in Evolutionary Biology. *Philosophy of Science*, Vol. 37, No. 1, pp. 1–15

Ayala, F.J. (1977). Teleological explanations. In T. Dobzhansky (ed.) *Evolution*, W.H. Freeman and Co., San Francisco, pp. 497–504.

Ayer, A.J. (1956). *The Problem of Knowledge*. London: MacMillan.

Bacon, F. (1620). *Novum Organum Scientiarum*. Venice: Gasparis Gerardis.

Bandea, C. (1983). A new theory on the origin and the nature of viruses. *Journal of Theoretical Biology, 105* (4), 591–602.

Banerjee, K., B. Huebner, & M. D. Hauser (2011). Intuitive moral judgments are robust across demographic variation in gender, education, politics, and religion: A large-scale web-based study. *Journal of Cognition and Culture*, 37, 151–187.

Barbieri, M. (2003). *The Organic Codes. An introduction to Semantic Biology*. Canbridge: Cambridge University Press.

Barnes, J. (1982). *The Presocratic Philosophers*, rev. ed., London and New York: Routledge.

Barrett, P. *et al.* (eds.) (1987). *Charles Darwin's Notebooks: 1836–1844*. Ithaca, NY: Cornell University Press.

Bassler, B.L., and Losick, R. (2006). Bacterially speaking. *Cell*. 2006 Apr 21;125(2):237–46.

Bealer, G. (1987). The Philosophical Limits of Scientific Essentialism. In Tomberlin, J. (ed.) *Philosophical Perspectives*, 1, 289–365. Ridgeview: Atascadero.

Bealer, George. (1998). Intuition and the Autonomy of Philosophy. In M. DePaul & W. Ramsey (eds.) *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Lanham, MD: Rowman & Littlefield. pp. 201–239.

Beatty, J. (1990). Teleology and the Relationship Between Biology and the Physical Sciences in the Nineteenth and Twentieth Centuries. In Durham and Purrington (eds.), *Some Truer Method: Reflections on the Heritage of Newton*, New York: Columbia University Press, pp. 113–144.

Beauvilliers, A.B. (1814). *L'art du Cuisinier*. Paris: Pilet.

Bechtel, W. (1986). Teleological functional analyses and the hierarchichal organization of nature. In N. Rescher, ed. *Current Issues in Teleology*. Lanham, MD: University Press of America.

Beckner, M. (1969). Function and teleology. *J. Hist. Biol.* 2,151–164.

Bedau, M. (1990). Against mentalism in teleology, *Amer. Phil. Quart.* 27(1), 61–70.

Bedau, M. (1991). Can biological teleology be naturalized? *J. Phil.* 88, 647–57.

Bedau, M. (1992a). Goal-directed systems and the good. *The Monist*. 75:34–49.

Bedau, M. (1992b). Where's the good in teleology? *Phil. and Phenomenol. Res*. 52(4),781–805.

Bedau, M. and Cleland, C., (eds.) (2010). *The Nature of Life: Classical and Contemporary Perspectives from Philosophy and Science.* Cambridge: Cambridge University Press.

Bedau, M.A. and Packard, N.H. (1996). Measurement of evolutionary activity, teleology, and life. In C. G. Langton, C. E. Taylor, J. D. Farmer, and S. Rasmussen, eds., *Artificial Life II, SFI Studies in the Sciences of Complexity, Vol. X*, Addison-Wesley, Redwood City, CA, pp. 431–483.

Beer, R.D. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, M. Mataric, J. Meyer, J. Pollack and S. Wilson (Eds.), *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 421–429). MIT Press.

Beer, R.D. and Williams, P.L. (2015). Information processing and dynamics in minimally cognitive agents. *Cognitive Science* 39:1–38.

Belousov, B.P. (1959) "Периодически действующая реакция и ее механизм" [Periodically acting reaction and its mechanism]. Сборник рефератов по радиационной медицине. 147: 145.

Benfey, O.T. (1963). *Classics in the Theory of Chemical Combination.* New York: Dover.

Berg, H.C. (2004). *E. coli in Motion.* New York: Springer-Verlag.

Berg, L.S. (1926). *Nomogenesis or evolution determined by law.* Translated from Russian by J.N. Rostovtsow. London: Constable and Co., Ltd.

Bergson, H. (1911). *Creative Evolution.* Translated from French by Arthur Mitchell, New York: Henry Holt.

Bernard, C. (1879). *Leçons sur les phénomènes de la vie communs aux animaux et aux végétaux.* Paris: Baillière.

Bernoulli, D. (1738) Danielis Bernoulli Joh. Fil. *Hydrodynamica, sive, De viribus et motibus fluidorum commentarii.* Argentorati: Sumptibus Johannis Reinholdi Dulseckeri, typis Joh. Henr. Deckeri, Typographi Basiliensis.

Bernoulli D. (1741 (1736)). De legibus quibusdam mechanicis… *Commentarii Academiae scientiarum imperialis Petropolitanae* 8: 99–127.

Bertalanffy, L. von (1956) General Systems Theory. In General Systems, Yearbook of the Society for General Systems Research, vol. 1, pp. 1–10.

Bertalanffy, L. von (1968). *General System Theory: Foundations, Development, Applications.* New York: George Braziller.

Bertenthal, B.I. (1993). Infants' perception of biomechanical motions: intrinsic image and knowledge-based constraints. In: Granrud C. (ed.) *Visual perception and cognition in infancy.* Hillsdale, NJ: Erlbaum, pp. 175–214.

Berthollet, C.L. (1803). *Essai de statique chimique*, Paris: Chez Firmin Didot.

Berzelius, J.J. (1827). *Lehrbuch der Chemie* (Textbook of Chemistry), vol. III. Dresden: Arnold.

Berzelius, J.J. (1836). Einige Ideen über bei der Bildung organischer Verbindungen in die lebenden Naturwirksame ober bisher nicht bemerkte Kraft. *Jahres-Berkcht über die Fortschritte der Chemie* 15: 237–45.

Bigelow, J. and Pargetter, R. (1987). Functions. *J. Phil.* 86(4), 181–196.

Birch, H. G., and Rabinowitz, H. S. (1951). The negative effect of previous experience on productive thinking. *Journal of Experimental Psychology* 41(2), 121–125.

Bitounis, D., Fanciullino, R., Iliadis, A. and Ciccolini, J. (2012). Optimizing druggability through liposomal formulations: New approaches to an old concept. *International Scholarly Research Network (ISRN) Pharmaceutics* Volume 2012, Article ID 738432.

Block, N. and Fodor, J. (1972) What psychological states are not. In *Philosophical Review* 81 (April) pp. 159–81.

Blum, T., Denig, A., Logashenko, I., de Rafael, E., Roberts, B.L., Teubner, T., Venanzoni, G. (2013) The Muon (g-2) Theory Value: Present and Future. *arXiv*:1311.2198.

Blumenbach, J.F. (1789). *Über den Bildungstrieb*. Gottingen: J.C. Dieterich.

Blumenbach, J.F. (1787). *Institutiones Physiologicae*. Göttingen: J.C. Dieterich.

Bohr, Niels (1933). Light and Life. *Nature* 131, 421–423.

Bohr, Niels (1976), The Correspondence Principle. In Rosenfeld, L. and Nielsen, J. Rud, eds., *Niels Bohr, Collected Works, Volume 3, (1918–1923)* Amsterdam: North-Holland.

Bollinger, R. R., Barbas, A.S., Bush, E.L., Lin, S.S., and Parker W. (2007). Biofilms in the large bowel suggest an apparent function of the human vermiform appendix. *J Theor. Biol.* 249(4): 826–31.

Boltzmann, L. (1872). Weitere studien über das Wärmegleichgewicht unter Gasmolekülen. *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien, mathematisch-naturwissenschaftliche Classe*, 66, pp. 275–370.

Boltzmann, L. (1896/1964) *Lectures on Gas Theory*. Brush, S.G., Translator. Berkeley, CA: University of California Press.

Boorse, C. (1976). Wright on functions. *Philosophical Review* 85: 70–86.

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science*, 44, 542–573.

Boorse, C. (2002). A rebuttal on functions. In Ariew, A., Cummins, R., and Perlman, M. (eds.) *Functions: New essays in the philosophy of psychology and biology*. Oxford University Press: Oxford.

Borchert T.H. and Waterman J.P. (2017). *The Book of Miracles*. Taschen Verlag.

Borges, W. and Stern, J. M. (2007). The rules of logic composition for the Bayesian epistemic e-values. *Logic Journal of the IGPL*, 15 (5–6), 401–420.

Bouchard, F. (2004). *Evolution, Fitness and the Struggle for Persistence.* Ph.D. Thesis, Philosophy Department, Duke University.

Boyd, R. (1999). Homeostasis, species and higher taxa. In Wilson, R. (ed.) *Species: New Interdisciplinary Essays.* Cambridge MA: MIT Press, pp. 141–186.

Boveri, T.H. (1904). *Ergebnisse über die Konstitution der chromatischen Substanz des Zelkerns*. Fisher, Jena.

Bragg J.G. and Chisholm S.W. (2008). Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. *PLoS ONE* 3(10): e3550.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology.* Cambridge, MA: MIT Press.

Braithwaite, R.B. (1953). *Scientific Explanation* Cambridge: Cambridge University Press.

Brandon, R.N. (1981). Biological Teleology: Questions And Explanations. *Stud. Hbt. Phil. Sci.*, Vol. 12, No. 2, pp. 91–105.

Brandon, R.N. (2011). A General Case for Functional Pluralism. In P. Huneman (ed.), *Functions: selection and mechanisms*, Synthese Library, Studies in Epistemology, Logic, Methodology, and Philosophy of Science, pp. 97–104.

Briggs, T.S. and Rauscher, W.C. (1973). An Oscillating Iodine Clock. *J. Chem. Educ.* 50: 496.

Brillat-Savarin, J.A. (1825). *The Physiology of Taste: Or Meditations on Transcendental Gastronomy.* Translated by M. F. K. Fisher (1949). New York: Heritage Press.

Brillouin, L. (1953). Negentropy Principle of Information. *J. of Applied Physics*, v. 24(9), pp. 1152–1163.

Brillouin, L. (1962) *Science and Information Theory.* New York: Academic Press.

Brooks, R. (1990). Challenges for Complete Creature Architectures. In R. A. Brooks (Ed.), *Proceedings of First International Conference on Simulation of Adaptive Behavior*. Cambridge MA: MIT Press (pp. 434–443).

Brooks, R. (1991) Intelligence Without Reason. *Proceedings of 12th Int. Joint Conf. on Artificial Intelligence*, Sydney, Australia, August 1991, pp. 569–595.

Brown, A. (1998). *The Science of Selfishness (book review of Unweaving the Rainbow)*. Salon. Salon Media Group.

Brown, R. (1970). The burden of proof. *American Philosophical Quarterly*, 7 (1), 74–82.

Brown, Robert (1828). "A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies." *Phil. Mag.* 4, 161–173.

Brusca, R.C. and Gilligan, M.R. (1983). Tongue replacement in a marine fish (Lutjanus guttatus) by a parasitic isopod (Crustacea: Isopoda). *Copeia* 3 (3): 813–816.

Buffon, G. (1749). *Histoire naturelle, générale et particulière.* Paris : De l'Imprimerie Royale.

Buller, D.J. (1998). Etiological Theories of Function: A Geographical Survey. *Biology and Philosophy* 13: 505–527.

Buller D.J. (ed.) (1999) *Function, Selection, and Design*. Albany, NY: State University of New York Press.

Buller, D.J. (2001). Function and Teleology. *Encyclopedia of Life Sciences*. New York: John Wiley and Sons.

Burt, A. and Trivers, R. (2006). *Genes in Conflict: the biology of selfish genetic elements.* Cambridge MA: Harvard University Press.

Butler, S. (1879). *Evolution, Old and New, or, The Theories of Buffon, Dr. Erasmus Darwin, and Lamarck, as Compared with that of Mr. Charles Darwin.* London: Hardwicke and Bogue.

Butler, S. (1880). *Unconscious Memory: A Comparison between the Theory of Dr. Ewald Hering…and "The Philosophy of the Unconscious" by Dr. Edward von Hartmann; With Translations from these Authors.* London: David Bogue.

Butt, H.J., Graf, K., and Kappl, M. (2006). *Physics and Chemistry of Interfaces.* Weinheim: Wiley-VCH. pp. 269–277.

Cairns-Smith, A.G. (1982). *Genetic Takeover and the Mineral Origins of Life.* Cambridge: Cambridge University Press.

Cairns-Smith, A.G. (1985). *Seven Clues to the Origin of Life.* Cambridge: Cambridge University Press.

Call, J., and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences* 12, 5:187–192.

Callicott, J.B. (1989). *In Defense of the Land Ethic: Essays in Environmental Philosophy.* Albany NY: SUNY Press.

Campaner, M. (2010). Reductionist and Antireductionist Stances in the Health Sciences. In Stadler, F. (ed.) *The Present Situation in the Philosophy of Science*. New York: Springer, pp. 205–218.

Campbell, J. (1985). An Organizational interpretation of evolution. In *Evolution at a Crossroads*, D. Depew and B.H Weber (eds.). Cambridge, MA: MIT Press.

Canfield, J. (1964). Teleological explanations in biology. *British Journal for the Philosophy of Science*, 14: 285–95.

Carlson, C.E. (2015). The Proton Radius Puzzle. *arXiv*:1502.05314v1.

Carnap, R. (1936). Testability and Meaning. *Philosophy of Science*. vol. 3, No. 4.

Carnot, S. (1824). *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance.* Paris: Bachelier.

Carter, M. (2016). *Helping Children and Adolescents Think about Death, Dying and Bereavement.* London: Jessica Kingsley Publishers.

Castaneda, H.N. (1984). Causes, causity, and energy. In P. French, T. Uehling, Jr., and H. Wettstein (eds.), *Midwest Studies in Philosophy IX*. Minneapolis: University of Minnesota Press, 17–27.

Chalmers, D.J. (1995). Facing up to the Problem of Consciousness.. *Journal of Consciousness Studies* 2: 200–219.

Chalmers, D.J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chambers, R.G. (1960). Shift of an Electron Interference Pattern by an Enclosed Magnetic Flux, *Phys. Rev. Lett.* 5, 3.

Charles, D. (1984). *Aristotle's Philosophy of Action*. London: Duckworth.

Charles, D. (1995). Teleological Causation in the Physics. In Lindsay Judson, (ed.) *Aristotle's 'Physics': A Collection of Essays*. Oxford: Clarendon, pp. 101–28.

Cheney, D. L. and Seyfarth, R. M. (1990). *How monkeys see the world: Inside the mind of another species*. University of Chicago Press, Chicago.

Cheney, D. L. and Seyfarth, R. M. (1992). Précis of How monkeys see the world: Inside the mind of another species. *Behav. Brain Sci.* 15,135–182.

Chiel, H.J., Beer, R.D. and Gallagher, J.C. (1999). Evolution and analysis of model CPGs for walking I. Dynamical modules. *J. Computational Neuroscience* 7:(2): 99–118.

Chisholm, R.M. (1957). *Perceiving: A Philosophical Study*. Ithaca, NY: Cornell University Press.

Christensen, W.D. (1996). A Complex Systems Theory of Teleology. *Biology and Philosophy* 11: 301–320.

Christensen, W.D. and Bickhard, M.H. (2002). The Process Dynamics of Normative Function. *The Monist*, Vol. 85, No. 1, The Philosophy of Biology, pp. 3–28

Churchland, Patricia (1986). *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press.

Cicero, M.T. (1967). *De Natura Deorum*. (ed. Warmington, E. H.) (trans. Rackham, H.) Cambridge, MA: Harvard University Press.

Clausius, R. (1857). Über die Art der Bewegung, welche wir Wärme nennen. *Annalen der Physik*, 176 (3): 353–379.

Cleland, C. and Chyba, C. (2002). Defining "Life". *Origins of Life and Evolution of the Biosphere* 32: 387–393, 2002.

Collins, H. and M. Kusch (1998). *The Shape of Actions: What Humans and Machines Can Do*. Cambridge, MA, MIT Press.

Cope, D. (1996). *Experiments in Musical Intelligence*. Madison, WI: A-R Editions.

Costanza, R. (2004). Value Theory and Energy. *Encyclopedia of Energy*, Volume 6.

Couper, A.S. (1858). On a New Chemical Theory, *Phil. Mag.* 16, 104.

Craver, C. (2013). Functions and Mechanisms: A Perspectivalist View. In Philippe Huneman (ed.), *Functions: Selection and Mechanisms*. Springer. pp. 133–158.

Crawford, Adair (1779). *Experiments and Observations on Animal Heat, and the Inflammation of Combustible Bodies*. London: Murray.

Crick, F. (1968). The origin of the genetic code. *J Mol Biol.* 1968 Dec; 38(3):367–79.

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5 month old infants. *Cognition*. 107(2): 705–17.

Csibra, G., Bíró, S., Koós, S., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, 27, 111–133.

Csibra, G., & Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental Science*, 1, 255-259.

Csibra, G., & Gergely, G. (2007). 'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, 124, 60–78.

Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbanck, M. (1999). Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition*, 72, 253–284.

Cummins, R. (1975/1984) Functional Analysis. *J. Phil.* 72,741.765. Reprinted with minor alterations in Sober (1984b), pp. 386-407.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge MA: MIT Press.

Cummins, R. (2002). Neo-teleology. In Ariew, A., Cummins, R., and Perlman, M. (eds.) *Functions: New essays in the philosophy of psychology and biology*. Oxford: Oxford University Press.

Damuth, J., & Heisler, I. L. (1988). Alternative formulations of multilevel selection. *Biology and Philosophy*, 3, 407–30.

Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association*, Vol. 60, No. 3 (Jan., 1987), pp. 441–458.

Davies, P.S. (2001). *Norms of Nature: Naturalism and the nature of functions*. Cambridge, MA: MIT Press.

Darden, L. and Cain, J.A. (1989). Selection type theories. *Philosophy of Science* 56, no. 1: 106–129.

Darwin, C. (1859). *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. London: J. Murray.

Darwin, C. (1860/1993). *The Correspondence of Charles Darwin, 8*. Cambridge: Cambridge University Press.

Darwin, C. (1861/1994). *The Correspondence of Charles Darwin, 9*. Cambridge: Cambridge University Press.

Darwin, C. (1868). *The Variation of Animals and Plants under Domestication*. London: John Murray.

Darwin, C. (1869). *On the Origin of Species, 5th Edition*. London: John Murray.

Darwin, C. (1896). *The Variation of Animals and Plants Under Domestication*. New York: D. Appleton and Company.

Darwin, C. (1959). *The Life and Letters of Charles Darwin*. (Francis Darwin, Ed.) New York: Basic Books.

Darwin, C. (1964). *On the Origin of Species: A Facsimile of the First Edition*, Harvard University Press, Cambridge, MA.

Darwin, F. and Seward, A. C., eds. (1903). *More letters of Charles Darwin. A record of his work in a series of hitherto unpublished letters*. London: John Murray.

Dausmann, K. H.; Glos, J.; Ganzhorn, J. U. & Heldmaier, G. (2004). Hibernation in a tropical primate. *Nature* 429 (6994): 825–826.

Dawkins, R., (1976). *The Selfish Gene*. Oxford: Oxford University Press.

Dawkins, R., (1978). Replicator selection and the extended phenotype. *Zeittschriftfür Tierpsychologie*, 47: 61–76.

Dawkins, R., (1982a). *The Extended Phenotype*, Oxford: Freeman.

Dawkins, R. (1982b). Replicators and Vehicles. In King's College Sociobiology Group (eds.), *Current Problems in Sociobiology*, Cambridge: Cambridge University Press, pp. 45–64.

Dawkins, R. (1983). Universal Darwinism. In D.S. Bendall (Ed.) *Evolution from molecules to man*. Cambridge UK: Cambridge University Press, pp. 403–425.

Dawkins, R. (1986). *The Blind Watchmaker*. New York: Norton.

Dawkins, R. (1995). *River Out of Eden.* New York: Basic Books.

Dawkins, R. (2005). *The Ancestor's Tale*. New York: Mariner Books.

Dawkins, R. (2014). *What Scientific Idea is Ready for Retirement? Essentialism.* Edge.org (https://edge.org/response-detail/25366).

Deacon, T. (2013). *Incomplete Nature: How Mind Emerged From Matter.* New York: W.W. Norton and Co.

DeLancey, C.S. (2006). Ontology and Teleofunctions: A Defense and Revision of the Systematic Account of Teleological Explanation. *Synthese*, Vol. 150, No. 1 (May, 2006), pp. 69–98.

Dennett, D.C. (1969). *Content and Consciousness.* Boston: Routledge & Kegan Paul.

Dennett, D. C. (1984). *Elbow room.* Cambridge, MA: MIT Press.

Dennett, D.C. (1987a). *The Intentional Stance*, Cambridge MA: MIT Press.

Dennett, D.C. (1987b). Intentional Systems in Cognitive Ethology: The "Panglossian Paradigm" Defended. In *The Intentional Stance,* Cambridge, MA: MIT Press, pp. 237–86.

Dennett, D.C. (1989). The Origins of Selves. *Cogito*, 3. 163-73.

Dennett, D.C. (1991). Real Patterns. *The Journal of Philosophy,* 87: 27–51.

Dennett, D. C. (1995). *Darwin's Dangerous Idea.* New York: Simon & Schuster.

Dennett, D.C. (1998). *Brainchildren: Essays on Designing Minds.* Cambridge, MA: MIT Press.

Dennett, D.C. (2003). *Freedom Evolves.* New York: Viking Press.

Dennett, D.C. (2005). "*The Selfish Gene* as a Philosophical Essay." In Grafen, A. and Ridley, M. (eds.) *Richard Dawkins: How a Scientist Changed the Way We Think*. Oxford: Oxford University Press.

Dennett, D.C. (2007). Philosophy as Naïve Anthropology: Comment on Bennett and Hacker. In Robinson, D. (ed.) *Neuroscience and Philosophy: Brain, Mind, and Language*. New York: Columbia University Press.

Dennett, D.C. (2013). Kinds of Things—Towards a Bestiary of the Manifest Image. In Ross, D., Ladyman, J. and Kincaid, H. (eds.) *Scientific Metaphysics*. Oxford: Oxford University Press, pp. 96–107.

Dennett, D.C. (2014). The Evolution of Reasons. In Bashour, B. and Muller, H.D. (eds.) *Contemporary Philosophical Naturalism and Its Implications*. Routledge, 2014, pp. 47–62.

Dennett, D.C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. W.W. Norton and Company.

Derham, W. (1713). *Physico-theology: Or, A Demonstration of the Being and Attributes of God*. Edinburgh: J. Murray.

Descartes, R. (1641/1901) Meditations on First Philosophy: In which the existence of God and the immortality of the soul are demonstrated. In: Manley DB, Taylor CS, Veitch (Translator) J, eds. Meditationes de prima philosophia, in qua Dei existentia et animæ immortalitas demonstratur; 1901.

Descartes, René (1983) [1644, with additional material from the French translation of 1647]. *Principia philosophiae* (Principles of Philosophy). Translation with explanatory notes by Valentine Rodger and Reese P. Miller (eds.). Dordrecht: Reidel.

Descartes, R. (1664). Treatise on Man. In *The Philosophical Writings of Descartes, Volume 1*, translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch (1985). Cambridge: Cambridge University Press.

Dirac, P.A.M. (1933). The Lagrangian in quantum mechanics. *Phys. Z. der Sowjetunion* 3: 64–71.

Dobell, C. (1932). *Antony van Leeuwenhoek and his "Little Animals"; being some account of the father of protozoology and bacteriology and his multifarious discoveries in these disciplines*. New York: Harcourt, Brace and Co.

Dobzhansky, T., Ayala, F.J., Stebbins, G.L., Valentine, J.W. (eds.) (1977). *Evolution*. San Francisco, CA: Freeman.

Doerr, A. (2014) *All the Light We Cannot See*. New York: Scribner.

Doren, H.A. van (2007). *Tailor-made carbohydrate surfactants? Systematic investigations into structure-property relationships of N-Acyl N-Alkyl 1-Amino-1-Deoxy-D-Glucitols, Carbohydrates as Organic Raw Materials III*, Wiley-VCH Verlag GmbH, pp. 255-272.

Dretske, F. (1986). Misrepresentation. In R.J. Bogdan (ed.) *Belief: form, content and function*, Oxford: Oxford University Press, pp. 17–36.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Driesch, H. (1908). *The Science and Philosophy of the Organism*. London: Adam and Charles Black.

Driesch, H. (1914). *The History and Theory of Vitalism*. London: MacMillan and Co.

Du Bois-Reymond, E. (1848). *Untersuchungen über thierische Elektricität. Vol. I-II.* Berlin: G. Reimer.

Ducasse, C.J. (1925). Explanation, Mechanism, and Teleology. *The Journal of Philosophy*, Vol. 22, No. 6 (Mar. 12, 1925), pp. 150–155.

Duhem, P. (1906). *The Aim and Structure of Physical Theory*. Trans. 1954. Princeton, New Jersey: Princeton University Press.

Dumortier, B. (1832). *Recherches sur la structure comparée et le développement des animaux et des végétaux.* Bruxelles: M. Hayez.

Duncker, K. (1945). On Problem-Solving (J. F. Dashiell, Eds. & L. S. Lees, Trans.). *Psychological Monographs*, 58 (5, Whole No. 270). Washington, D.C.: The American Psychological Association, Inc. (Original work published 1935).

Eddington, A.S. (1928) *The Nature of the Physical World.* New York: The MacMillan Company.

Ehring, D. (1997). *Causation and Persistence.* Oxford: Oxford University Press.
Eibl-Eibesfeldt, I. (1975). *Ethology: The biology of behavior, Second Edition.* New York: Holt, Rinehart, and Winston.

Eigen, M. (1971). Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58: 465–523.

Eigen, M. (1983). Self Replicaton and Molecular Evolution. In D.S. Bendall (Ed.) *Evolution from molecules to man.* Cambridge, UK: Cambridge University Press, pp. 105–131.

Eigen, M. and Schuster, P. (1977). The Hypercycle: A Principle of Natural Self-Organization. Part A: Emergence of the Hypercycle. *Die Naturwissenschaften*, 64:541–565.

Eigen, M. and Schuster, P. (1979). *The Hypercycle: A Principle of Natural Self-Organization.* Berlin: Springer-Verlag.

Eimer, G.H.T. (1890). *Organic evolution as the result of the inheritance of acquired characters according to the laws of organic growth.* (Trans. by J.T. Cunningham). London: MacMillan and Co.

Eimer, G.H.T. (1897). *On Orthogenesis and the Impotence of Natural Selection in Species Formation.* Chicago: The Open Court Publishing Co.

Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik* (in German) 322 (8): 549–560.

Einstein, A. (1931). In Living Philosophies: A Series of Intimate Credos. Henry Goddard Leach (ed.). New York: Simon and Schuster.

Ekbia, H. (2008). Artificial dreams: The quest for non-biological intelligence. Cambridge, UK: Cambridge University Press.

Elgin, M. (2010). Reductionism in Biology: an Example of Biochemistry. In F. Stadler, D. Dieks, W. Gonzales, S. Hartmann, T. Uebel & M. Weber (eds.) *The Present Situation in the Philosophy of Science,* Springer, pp 195–203.

Elster, J. (1979). *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Enç, B. (1979) Function Attributions and Functional Explanations. *Philosophy of Science*, Vol. 46, No. 3 (Sep., 1979), pp. 343–365.

Enç, B. and Adams, F. (1992). Functions and goal directedness. *Phil. Sci.* 59(4),635–654.

Endoh, Y. (2014). Quine relay: An uroboros program with 100+ programming languages. Code downloaded from https://github.com/mame/quine-relay

Endoh, Y. (2015). あなたの知らない超絶技巧プログラミングの世界 (The World of Obfuscated, Esoteric, Artistic Programming). Tokyo: 技術評論社.

Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, 14, 219–50.

Farmer, D. & Belin, A. (1992). Artificial life: The coming evolution. In C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen (Eds.), *Artificial life II (Santa Fe Institute studies in the sciences of complexity, proceedings vol. X).* Redwood City, CA: Addison-Wesley, pp. 815–840.

Feynman, Richard P. (1942). The Principle of Least Action in Quantum Mechanics, Ph.D. Dissertation, Princeton University. Reprinted as Brown, L.M. (ed.) *Feynman's Thesis: a New Approach to Quantum Theory*, 2005, World Scientific Publishers.

Feynman, R. P., Leighton, R. B., & Sands, M. L. (1963). *The Feynman lectures on physics*. Reading, MA: Addison-Wesley Pub. Co.

Fine, A. (1986). *The Shaky Game: Einstein, Realism, and the Quantum Theory*. Chicago: University Press.

Frank, S.A. (1996). The Design of Natural and Artificial Adaptive Systems. In Rose and Lauder (eds.) *Adaptation.* San Diego: Academic Press.

French, R. (1995). *The Subtlety of Sameness: A Theory and Computer Model of Analogy Making.* Cambridge, MA: MIT Press.

Fodor, J.A. (1974). Special Sciences. In Fodor, J.A. (ed.) *Representations: Philosophical Essays on the Foundations of Cognitive Science*, 1981, Cambridge, MA: MIT Press, pp. 127–145.

Fodor, J.A. (1981). The Present Status of the Innateness Controversy. In *Representations: Philosophical Essays on the Foundations of Cognitive Science.* Cambridge MA: MIT Press, pp. 257–316.

Forterre, P. (2010). Defining Life: The Virus Viewpoint *Origins of Life and Evolution of Biospheres, 40* (2), 151–160.

Frank, M.C. and Ramscar, M. (2003). How do presentation and context influence representation for functional fixedness tasks? *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, Boston, MA.

Frank, S.A. (1996). The Design of Natural and Artificial Adaptive Systems. In Rose and Lauder (eds.) *Adaptation*. San Diego: Academic Press.

Fukuda, H., and Ueda, K. (2010). Interaction with a Moving Object Affects One's Perception of Its Animacy. *International Journal of Social Robotics*, Volume 2, Number 2, 187–193.

Gaisford, T. (1850). *Ioannos Stobaie: Eclogarum Physicarum Et Ethicarum, Libri Duo.* Oxford: Academic Press.

Galilei, G. (*ca.* 1590). De motu. In Drabkin, I. E., In Drake, S., & Galilei, G. (1590/1960). *On motion, and On mechanics: Comprising De motu*. Madison: University of Wisconsin Press.

Galilei, G. (1616) Discorso Sul Flusso E Il Reflusso Del Mare. Letter to Cardinal Alessandro Orsini. In *The Galileo Affair: A Documentary History*. Berkeley: University of California Press, 1989.

Galilei, G. (1623) *Il Saggiatore* (The Assayer). As Drake, S. (trans.) *Discoveries and Opinions of Galileo*, 1957. New York: Anchor Books.

Gánti, T. (1971). *The Principle of Life* (in Hungarian). Budapest: Gondolat. Translated from the Hungarian as *The Principles of Life*, 2003, Oxford: Oxford University Press.

Gánti, T. (1997). Biogenesis itself. *J. Theor. Biol.* 187, 583–593.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, 59, 154–179.

García-Ruiz, J.M., Villasuso, R., Ayora, C., Canals, A., and Otálora, F. (2007). Formation of natural gypsum megacrystals in Naica, Mexico. *Geology* v. 35 no. 4 p. 327–330.

Garson, J. (2012). Selected effects and causal role functions in the brain: the case for an etiological approach to neuroscience. *Biol Philos* 26:547–565

Geach, P. (1972), Some Problems about Time. In *Logic Matters*, Berkeley, CA: University of California Press, pp. 302–318.

Gergely, G., & Csibra, G. (1997). Teleological reasoning in infancy: The infant's naive theory of rational action. A reply to Premack & Premack. *Cognition*, 63, 227–233.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The one-year-old's naïve theory of rational action. *Trends in Cognitive Sciences*, 7, 287–292.

Gergely, G., Na´dasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56, 165–193.

Gershenson (2010). The World as Evolving Information. *arXiv*: 0704.0304v3.

Gettier E.L. (1963). Is justified true belief knowledge? *Analysis*, Volume 23, Issue 6, 1 June 1963, Pages 121–123.
Gibbons , A. (2012). The Ultimate Sacrifice. *Science.* Vol. 336, Issue 6083, pp. 834–837.

Gibbs, J.W. (1873). A Method of Geometrical Representation of the Thermodynamic Properties of Substances by Means of Surfaces. *Transactions of the Connecticut Academy of Arts and Sciences 2*, Dec. 1873, pp. 382–404.

Gilbert, S. and Sarkar, S. (1998). Embracing Complexity: Organicism for the 21st Century. *Developmental Dynamics* 219: 1–9.

Gilbert, W. (1986). The RNA world. *Nature*, p. 618 v. 319.

Gilmore, G.W. (1919). *Animism or thought currents of primitive peoples*. Boston, MA: Marshall Jones Co.

Glover, J. (Interviewee). (2011, Oct 9). Philosophy Bites: Systems of Belief [Audio podcast]. Retrieved from http://philosophybites.libsyn.com/jonathan-glover-on-systems-of-belief.

Gödel, K. (1931) "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I", Monatshefte für Mathematik und Physik, 38: 173–198. Reprinted in Gödel 1986, pp. 144–195.

Godfrey-Smith, P. (1993). Functions: Consensus without unity. *Pacific Philosophical Quarterly*, 74, 196–208.

Godfrey-Smith, P. (1994). A modern history theory of functions. *Nous* Volume 28, Issue 3 (Sep., 1994), 344-362.

Godfrey-Smith, P. (2009) *Darwinian Populations and Natural Selection.* Oxford: Oxford University Press.

Goldschmidt, R. (1940/1982). *The Material Basis of Evolution.* New Haven CT: Yale University Press.

Goldschmidt, V. (1952). Geochemical aspects of the origin of complex organic molecules on the Earth, as precursors to organic life. *New Biology.* 12: 97–105.

Goldstein, D.S. (2006). *Adrenaline and the Inner World: An Introduction to Scientific Integrative Medicine.* Baltimore MD: Johns Hopkins University Press.

Goldstein, K. (1939/1995). *The Organism: A Holistic Approach to Biology Derived from Pathological Data in Man.* New York: Zone Books.

Gould, S.J. and Lewontin, R. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. Roy. Soc. London.* Reprinted in Sober 1984b, pp. 252–270.

Gould, S.J. (1982). *The Panda's Thumb: More Reflections in Natural History.* New York: Norton.

Gould, S.J., and Vrba, E.S. (1982). Exaptation—A Missing Term in the Science of Form. *Paleobiology*. 8:4–15.

Gray, A. (1874). Scientific Worthies: Charles Darwin. *Nature* 10, pp. 81.

Gray, A. (1963a). Natural Selection not Inconsistent with Natural Theology. In *Darwiniana* (Cambridge, MA: The Belknap Press of Harvard University), 121–2.

Griffiths, P. (1992). Adaptive explanation and the concept of a vestige. In P. Griffiths (ed.) *Trees of Life*, Netherlands: Kluwer, pp. 111–131.

Griffiths, P.E. (1993). Functional analysis and proper functions. *Brit. J. Phil. Sci.* 44.

Griffiths, P.E. (1999). Squaring the Circle: Natural Kinds with Historical Essences. In R. Wilson (ed.) *Species: new interdisciplinary essays*, Cambridge, MA: MIT Press, 209–228.

Griffiths, P.E. (2009). In what sense does 'nothing in biology make sense except in the light of evolution'? *Acta Biotheoretica*, 57: 11-32.

Griffiths, P.E. and Gray, R.D. (1994). Developmental Systems and Evolutionary Explanation. *Journal of Philosophy*, XCI (6): 277–304.

Grinstead, C.M. and Snell, J.L. (2006). *Introduction to Probability*. 2nd Edition. American Mathematical Society.

Guo, D.L., Xia, M.X., Wei, X., Chang, W.J., Liu, Y., and Wang Z.Q. (2008). Anatomical traits associated with absorption and mycorrhizal colonization are linked to root branch order in twenty-three Chinese temperate tree species. *New Phytologist* 180: 673–683.

Haacke, W. (1893). *Gestaltung und Vererbung.* Leipzig: Weigel.

Haig, D. (1997). The Social Gene. In Krebs, J. R. and Davies, N. B. (eds.) *Behavioural Ecology*. UK: Blackwell Scientific, pp. 284–304.

Haig, D. (2013). Proximate and Ultimate Causes: How Come? and What For? *Biol Philos* 28 (5) (April 3): 781–786.

Haldane, J.S. (1931). *The philosophical basis of biology*. London: Hodder and Stoughton.

Hall, R.J. (1990). Does representational content arise from biological function? PSA vol. 1, pp. 193-199.

Haller, A. von (1766). *Elementa Physiologiae corporis humani.* Bern: Bousquet, D'Arnay, and Grasset.

Hamilton, W.D. (1963). The evolution of altruistic behaviour. *American Naturalist*. 97: 354–6.

Hamilton, W.D. (1964). The genetical evolution of social behaviour, I. *J. Theor. Biol.* 7 (1): 1–16.

Hamilton, W.D. (1964). The genetical evolution of social behaviour, II. *J. Theor. Biol.* 7 (1): 17–52.

Hanke, David (2004). Teleology: The explanation that bedevils biology. In John Cornwell (ed.). *Explanations: Styles of explanation in science*. Oxford & New York: Oxford University Press. pp. 143–155.

Hardcastle, V.G. (2002). On the normativity of functions. In Ariew, A., Cummins, R., and Perlman, M. (eds.) *Functions: New essays in the philosophy of psychology and biology.* Oxford: Oxford University Press.

Hare, B., Call, J., and Tomasello, M. (2001). Do Chimpanzees Know What Conspecifics Know? *Animal Behaviour,* 61:139–151.

Häring M, Vestergaard G, Rachel R, Chen L, Garrett RA, Prangishvili D (2005). Virology: independent virus development outside a host. *Nature* 436:1101–1102.

Harvey, William (1628/1889). *On the Motion of the Heart and Blood in Animals*. London: George Bell and Sons.

Hauser, M.D. (2006). *Moral minds: How nature designed our universal sense of right and wrong.* New York, NY: Ecco/HarperCollins Publishers.

Hauser, M.D., Cushman, F.A., Young, L., Jin, R., & Mikhail, J.M. (2007). A dissociation between moral judgment and justification. *Mind and Language*, 22, 1–21.

Hayes, W. (2007). Is the outer Solar System chaotic?. *Nature Physics* 3 (10): 689–691.

Heath T.L. (trans.) (1897) *The Sand Reckoner of Archimedes.* Cambridge, UK: Cambridge University Press.

Hebb, D. (1949). *The organization of behavior: A neuropsychological theory.* New York: John Wiley and Sons.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behaviour. *American Journal of Psychology*, 57, 243–249.

Helmholtz, H. von (1845). 'Wärme, physiologisch', In Helmholtz, Hermann. *Wissenschaftliche Abhandlungen.* 2nd vol. Leipzig: Barth, 1883.

Helmholtz, H. von (1882). On the Thermodynamics of Chemical Processes. In *Physical Memoirs Selected and Translated from Foreign Sources*, 1: 43-97. London: Physical Society of London, Taylor and Francis, 1888.

Helmisaari, H.S., Derome, J., Nöjd, P., and Kukkola, M. (2007). Fine root biomass in relation to site and stand characteristics in Norway spruce and Scots pine stands. *Tree Physiology* 27, 1493–1504.

Hempel, Carl Gustav (1959/1965). The Logic of Functional Analysis. In Llewellyn Gross, ed. *Symposium on Sociological Theory*. New York: Harper and Row. 271–307. Reprinted in Hempel. *Aspects of Scientific Explanation*. New York: Free Press. 297–330.

Herodotus (2007). *The Landmark Herodotus: The Histories*. New York: Pantheon Books.

Hicks, R.D. (ed.) (1925/1972). *Diogenes Laertius: Lives of eminent philosophers* (Vols. 1-2). Cambridge, MA: Harvard University Press, Loeb Classical Library.

Hinde, R.A. (1975). The concept of function. In Baerends *et al.* (eds.) *Function and Evolution of Behavior*. Oxford: Oxford University Press, pp. 3–15.

Hofstadter, D.R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. NY: Basic Books.

Hofstadter, D.R. (1980). Reductionism and Religion. *Behavioral and Brain Sciences,* 3 (3): 433–434.

Hofstadter, D.R. (1985). *Metamagical Themas: Questing for the Essence of Mind and Pattern* NY: Basic Books.

Hofstadter, D.R. (2007). *I Am a Strange Loop*. New York: Basic Books.

Hofstadter, D.R. and the Fluid Analogies Research Group. (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books.

Hofstadter, D.R. and Sander, E. (2013). *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. NY: Basic Books.

Holland, J.H. (1995). *Hidden Order: How Adaptation Builds Complexity*. Reading MA: Perseus Books.

Holland, J.H. (1998). *Emergence: From Chaos to Order*. New York: Basic Books.

Hooke, R. (1665). *Micrographia, or some physiological descriptions of minute bodies made by magnifying glasses, with observations and inquiries thereupon*. London: Martyn and Allestry.

Hull, D. (1965). The effect of essentialism on taxonomy: two thousand years of stasis. *British Journal for the Philosophy of Science*. 15: 314–326, 16: 1–18.

Hull, D.L. (1973). *Darwin and His Critics: The Reception of Darwin's Theory of Evolution by the Scientific Community*. Cambridge, MA: Harvard University

Hull, D.L. (1974). *Philosophy of Biological Science*. Englewood Cliffs, NJ: Prentice Hall.

Hull, D.L. (1978). A matter of individuality, *Philosophy of Science* 45: 335–360.

Hull, D.L. (1980). Individuality and Selection, *Annual Review of Ecology and Systematics*, 11: 311–332.

Hull, D.L. (1988). Interactors versus vehicles. In Plotkin, H.C. (ed.) *The role of behavior in evolution*. Cambridge, MA: MIT Press, pp. 19–50.

Humberstone, I. L. (1996), Intrinsic/Extrinsic, *Synthese*, 108: 205–67.

Hume, David (1739). *A Treatise of Human Nature*. London: John Noon.

Hume, David (1748). *An Enquiry Concerning Human Understanding*. London: A. Millar.

Hume, David (1779). *Dialogues Concerning Natural Religion* (2nd ed.). London.

Hurley, M.M., Dennett, D.C., & Adams Jr., R.B. (2011). *Inside Jokes: Using Humor to Reverse-Engineer the Mind*. Cambridge, MA: MIT Press.

Hurley, M.M., Dennett, D.C., & Adams Jr., R.B. (2013, August 8). The Complicated Search for the Origins of Humor. *Huffington Post*. Retrieved from https://www.huffingtonpost.com

Huxley, T.H. (1863). *Evidence as to Man's Place in Nature*. London: Williams and Norgate.

Inwagen, P. van and Sullivan, M., (2018) Metaphysics. In Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). Retrieved from https://plato.stanford.edu/archives/spr2018/entries/metaphysics/

Ireland, J.D. (trans.) (1997) *The Udana and the Itivuttaka: Two Classics from the Pali Canon*. Sri Lanka: Bahirawakanda, Kandy.

Iversen,C.M., Sloan V.L., Sullivan, P.F., Euskirchen, E.S., McGuire, A.D., Norby, R.J., Walker, A.P., Warren, J.M., and Wullschleger, S.D. (2015). The Unseen Iceberg: Plant Roots in Arctic Tundra. *New Phytologist*. Volume 205, Issue 1, pages 34–58, January 2015.

Jackendoff, R.S. (1983). *Semantics and Cognition*, Cambridge, MA: MIT Press.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Jacobs, L.F., & Liman, E.R. (1991). Grey squirrels remember the locations of buried nuts. *Anim. Behav*. 41, 103–110.

Ji, Q. and Ji, S. (1996). "On discovery of the earliest bird fossil in China (*Sinosauropteryx* gen. nov.) and the origin of birds". *Chinese Geology*. Beijing: Chinese Geological Museum 10 (233): 30–33.

Juarrero, A. (1999). *Dynamics in Action: Intentional Behavior as a Complex System*. Cambridge, MA: MIT Press.

Kant, Immanuel (1787). *Critique of Pure Reason*. Translated from the German as Guyer, P. and Wood, A., (eds.), 1998, Cambridge: Cambridge University Press.

Kant, Immanuel (1790) *Critique of the Power of Judgment*. Translated from the German as Guyer, P. and Matthews, E. (eds.), 2000, Cambridge: Cambridge University Press.

Kauffman, S.A. (1971a). Articulation of parts explanations in biology. *Boston studies in the philosophy of science*, ed. R. S. Cohen and R. C. Buck, VIII (Dordrecht), 257-72.

Kauffman, S.A. (1971b). Cellular Homeostasis, Epigenesis, and Replication in Randomly Aggregated Macromolecular Systems. *Journal of Cybernetics* 1: 71–96.

Kauffman, S.A. (2000). *Investigations*. Oxford: Oxford University Press.

Keil, F.C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. A. Hirschfeld & S. A. Gelman (eds.), *Mapping the mind: Domain specificity in cognition and culture*. New York: Cambridge University Press, pp. 234-254.

Keil, F.C. (1995). The growth of causal understanding of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate*. Oxford: Oxford University Press, pp. 234–262.

Kekulé, A. (1858). Über die Konstitution und Metamorphosen der chemischen Verbindungen und über die chemische Natur des Kohlenstoffs. Ann. Chem. 106:129-152. Reprinted in English Translation in Benfey, O.T. *Classics in the Theory of Chemical Combination*. New York: Dover Publications, 1963, pp. 109–131.

Keller, E.F. (2007). The disappearance of function from "self-organizing systems". In Boogerd, F., Bruggeman, F., Hofmeyr, J.-H., and Westerhoff, H.V. (eds.), *Systems biology: philosophical foundations*. Amsterdam: Elsevier, pp. 303–317.

Kellogg, V.L. (1906). Is there determinate variation? *Science* 24: 621-628.

Kellogg, V.L. (1907). *Darwinism To-Day: a discussion of present-day scientific criticism of the Darwinian selection theories, together with a brief account of the principal other proposed auxiliary and alternative theories of species-forming*. New York: Henry Holt.

Kiedrowski, G. von (1986). A Self-Replicating Hexadeoxynucleotide, *Angw. Chem. Int. Ed. Eng.* 25, 932–934.

Kim, J. (1996). *Philosophy of Mind*. Boulder CO: Westview Press.

King, L. (1964). Stahl and Hoffmann: A Study in Eighteenth Century Animism. *Journal of the History of Medicine and Allied Sciences*, Volume XIX, Issue 2, 1 April 1964, Pages 118–130.

Kirschner, M., Gerhart, J., and Mitchison, T. (2000) Molecular "Vitalism". *Cell*, 100: p. 87.

Kistler, M. (1998). Reducing causality to transmission. *Erkenntnis*, 48, 1–24.

Kitcher, P. (1993). Function and Design. *Midwest Studies in Philosophy*, 19, 379-397. Mineapolis MN: University of Minnesota Press.

Knobe, J. & Nichols S. (eds.) (2008). *Experimental Philosophy*. Oxford: Oxford University Press.

Koffka, K. (1935). *Principles of Gestalt Psychology*. New York: Harcourt, Brace.

Kondepudi, D. and Prigogine, I. (2015). *Modern Thermodynamics: From Heat Engines to Dissipative Structures, Second Edition.* New York: John Wiley and Sons.

Kosman, Aryeh (1969). Aristotle's Definition of Motion, *Phronesis* 14 (1): 40–62.

Krebs, J.R. and Davies, N.B. (1997). *Behavioural Ecology: An Evolutionary Approach*. New York: Wiley-Blackwell.

Kripke, S. (1972). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kuhn, T. (1990). Dubbing and Re-Dubbing: The Vunerability of Rigid Designation. In Savage, W. (ed.) *Scientific Theories*. Minneapolis MN: University of Minnesota Press, pp. 293–318.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind.* Chicago: University of Chicago Press.

Lambert, F. (2002). Disorder—A Cracked Crutch For Supporting Entropy Discussions. *J. Chem. Educ.* 79, 187-192.

Lane, N. (2009). *Life Ascending: The Ten Great Inventions of Evolution*. New York: W.W. Norton.

Laplace, P.S. (1796). *Exposition Du Système Du Monde*. Paris: Fayard.

Laplace, P.S. (1814) *A Philosophical Essay on Probabilities*, translated into English from the original French 6th ed. by Truscott, F.W. and Emory, F.L., New York: Dover Publications, 1951.

Laubichler, M.D. (1999). A semiotic perspective on biological objects and biological functions. *Semiotica* 127-1/4, pp. 415–431.

Laurence, S. and Margolis, E. (2003). Concepts and conceptual analysis. *Philosophy and Phenomenological Research*. 67 (2): 253–282.

Lavoisier, A.L., and Laplace, P.S. (1780) "Mémoire sur la. Chaleur". *Mém. Acad. Roy. Sci.*, 355-408. Also in *Oeuvres de Lavoisier, publiées par les soins du Ministre de l'Instruction Publique, 2*. (1862), 283-333. Also in a joint French and English translation, H. Guerlac, Ed., *Memoir on Heat*, New York, 1982.

Lees, J.P., *et al.* (BaBar Collaboration) (1970). Evidence for an excess of $B \rightarrow D(*)\tau-\tau\nu$ decays. *Physical Review Letters,* 109(10).

Leeuwen N. van (2013). Imagination in Action. In *The Routledge Handbook of Philosophy of Imagination*, (ed. Amy Kind). London: Routledge.

Leff, H.S. (1996). Thermodynamic entropy: The spreading and sharing of energy. *Am. J. Phys.* 64: 1261–71.

Lehman, H. (1965). Functional explanation in biology, *Phil. Sci.*, 3, 1–20.

Lehn, J.M. (2002). Toward complex matter: Supramolecular chemistry and self-organization. *PNAS* 99 (8) 4763–4768.

Leibniz, G.W. (1673). Methodus Tangentium Inversa, Seu de Functionibus. in *Mathematischen Schriften*, herausgegeben vom Leibniz Archiv der Niedersächischen Landesbibliothek, Hannover, dritter Band (1672–1676), pp. 193-201.

Lennox, J. (1993). Darwin was a Teleologist. *Biology and Philosophy* 8: 409–421.

Lewens, T. (2004). *Organisms and Artifacts: Design in Nature and Elsewhere.* Cambridge MA: MIT Press.

Lewes, G.H. (1875). *Problems of Life and Mind* (First Series), 2, London: Trübner.

Lewis, D. (1969). Review of Art, Mind, and Religion. *Journal of Philosophy* 66: 23–35.

Lewis, D. (1986). Causation. In *Philosophical Papers, Volume II.* New York: Oxford, pp. 172-213.

Lewis, D. (1994). Reduction of Mind. In D. Lewis (ed.) *Papers in Metaphysics and Epistemology*, 1999, Cambridge: Cambridge University Press, pp. 291–324.

Lewontin, R.C. (1970). The units of selection. *Annual Review of Ecology and Systematics*, 1: 1–18.

Lewontin, R.C. (1993). *Biology as Ideology: The Doctrine of DNA.* New York: Harper Collins.

Liebig, J. von (1843/2002). Animal chemistry, or organic chemistry in its application to physiology and pathology. Chestnut Hill, MA: Adamant Media Corporation. Facsimile reprint of 1843 edition by Taylor and Walton, London.

Liebig, J. von (1844). *Familiar Letters on Chemistry, and its Relation to Commerce, Physiology, and Agriculture.* Gardner, J. (ed.) London: Taylor and Walton.

Lindström, B., and Pettersson, L. (2003) A Brief History of Catalysis. *CATTECH* 7: 130–138.

Lorenz, K. (1966). *On Aggression.* Latzke, M. (trans.) London: Methuen.

Lorenz, K. (1981). *The Foundations of Ethology,* New York: Springer-Verlag.

Lotka, A.J. (1910) Contributions to the Theory of Periodic Reactions. *J. Phys. Chem.* 14, 3, 271–274.

Lotze, R.H. (1842). Leben und Lebenskraft. In Wagner, R. (ed.), *Handwörterbuch der Physiologie*, Band 1. Braunschweig: F. Bieweg und Sohn, pp. 9–58.

Lovejoy, A.O. (1936/2009). *The Great Chain of Being: The Study of the History of an Idea.* New Brunswick, NJ: Transaction Publishers.

Lykken, J.D. (2010). Beyond the Standard Model. *CERN Yellow Report.* CERN. pp. 101–109.

Mace, C.A. (1935). Mechanical and teleological causation, *Proc. Aris. Soc.*, Supp., 14 (1935); reprinted in H. Feigl and W. Sellars, (eds.) *Readings in philosophical analysis.* 1949, New York: Appleton-Century-Crofts, pp. 534-39.

MacIntyre, A. (1981). *After Virtue: A Study in Moral Philosophy*. Notre Dame, IN: University of Notre Dame Press.

Mackie, J.L. (1974). *The Cement of the Universe.* Oxford: Oxford University Press.

MacLeod, R.B. (1957). Teleology and Theory of Human Behavior. *Science* 125: 477.

Madruga, M. R., Pereira, C. A. B., & Stern, J. M. (2003). Bayesian evidence test for precise hypotheses. *Journal of Statistical Planning and Inference*, 117 (2), 185–198.

Maier, N. R. F. (1930). Reasoning in humans I: On direction. *Journal of Comparative Psychology*, 10(2), 115–143.

Maier, N. R. F. (1931). Reasoning in humans: II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, 12(2), 181–194.

Maimonides (trans.: Pines, S.) (1963) *The Guide of the Perplexed.* Chicago: University of Chicago Press.

Malthus, T.R. (1826). *An Essay on the Principle of Population.* London: J. Johnson in St. Paul's Churchyard.

Manning, R.N. (1997). Biological Function, Selection, and Reduction. *British Journal for the Philosophy of Science*, 48:69–82.

Marchi, S., Bottke, W.F., Elkins-Tanton, L.T., Bierhaus, M., Wuennemann, K., Morbidelli, A. & Kring, D.A. (2014). Widespread mixing and burial of Earth's Hadean crust by asteroid impacts. *Nature* 511, 578–582.

Marx, K. (1861). Marx to Ferdinand Lassalle. In Berlin, 16 January 1861. *Marx & Engels Internet Archive: Letters*. https://www.marxistsfr.org/archive/marx//works/1861/letters/61_01_16.htm. Also in F. Lassalle. *Nachgelassene Briefe und Schriften*, Stuttgart, 1922.

Matthen, M. (1988). Biological Functions and Perceptual Content. *The Journal of Philosophy*, Vol. 85, No. 1 (Jan., 1988), pp. 5–27.

Matthen, M. (1997). Teleology and the product analogy. *Australasian Journal of Philosophy*, 75:1, 21–37.

Matthen, M. and Levy, E. (1984). Teleology, Error, and the Human Immune System. *The Journal of Philosophy*, Vol. 81, No. 7 (Jul., 1984), pp. 351–372.

Maturana, H.R., and Varela, F.J., Eds. (1973/1980). *Autopoiesis and Cognition: The Realization of the Living*, Dordrecht: D. Reidel Publishing Company.

Maupertuis, P.L.M. de (1745). *Venus Physique*. Trans. Simone Brangier Boas (1966). New York and London: Johnson Reprint Corp.

Maupertuis, P.L.M. de (1746). Les Loix du Mouvement et du Repos Déduites d'un Principe Metaphysique. *Mémoires de l'académie des sciences de Berlin* 1746, p. 267–294.

Maxwell, J.C. (1860). Illustrations of the dynamical theory of gases. Part I. On the motions and collisions of perfectly elastic spheres. *Philosophical Magazine*, 4th series, 19 : 19–32.

Maxwell, J.C. (1860). Illustrations of the dynamical theory of gases. Part II. On the process of diffusion of two or more kinds of moving particles among one another. *Philosophical Magazine*, 4th series, 20 : 21–37.

Maxwell, J.C. (1868). On Governors. from *the Proceedings of the Royal Society*, No.100, 1868.

Maxwell, J.C. (1873). Molecules. *Nature*, 417 (6892): 903.

Maynard Smith, J. (1964). Group Selection and Kin Selection. *Nature*. 201 (4924): 1145–1147.

Maynard Smith, J. (1998). The Units of Selection. *Novartis Found Symp*. 213: 203–11.

Maynard Smith, J. and Szathmáry, E. (1995). *The Major Transitions in Evolution*. Oxford: Oxford University Press.

Mayr, E. (1961). Cause and effect in biology. *Science* 134:1501–1506. Reprinted in Lerner, D. (ed.) *Cause and effect*. New York: Free Press, pp. 33–50.

Mayr, E. (1974/1988). *The multiple meanings of teleological*. Reprinted with a new postscript in Mayr (1988), pp. 38-66.

Mayr, E. (1988). *Toward a New Philosophy of Biology: Observations of an Evolutionist*. Cambridge MA: Harvard University Press.

Mayr, E. (1992). The idea of teleology. *Journal of the History of Ideas*, 53, 117–135.

Mayr, E. (1997). What is the meaning of "life"? in *This is Biology: The Science of the Living World*. Cambridge MA: Harvard University Press.

Mayr, E. (2002). The autonomy of biology. *Ludus vitalis*. vol. XII, num. 21.

McCorduck, P. (1991). *Aaron's Code*. New York: W.H. Freeman.

McLaughlin, P. (2001). *What Functions Explain: Functional Explanation and Self-Reproducing Systems*. Cambridge: Cambridge University Press.

McShea, D. W. (2012). Upper directed systems: a new approach to teleology in biology. *Biol Philos* 27:663–684.

McShea, D.W. and Brandon, R.N. (2010). *Biology's First Law: The Tendency for Diversity and Complexity to Increase in Evolutionary Systems.* Chicago: University of Chicago Press.

Medawar, P.B. (1952). An Unsolved Problem in Biology. London: Lewis. Reprinted in Medawar, P.B. (ed.), l981.*The Uniqueness of the Individual*. New York: Dover.

Melander, P. (1997). *Analyzing Functions: An Essay on a Fundamental Notion in Biology.* Stockholm: Almkvist & Wiksell International.

Meltzoff, A. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31, 838–850.

Menabrea, Luigi Federico; Lovelace, Ada (1843). Sketch of the Analytical Engine invented by Charles Babbage . . . with notes by the translator. Translated by Ada Lovelace. In Richard Taylor. *Scientific Memoirs*. 3. London: Richard and John E. Taylor. pp. 666–731.

Menzies, P. (2014) Counterfactual Theories of Causation. In Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Downloaded from http://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/

Mesmer, F.A. (1814). *Mesmerism; or, the System of Mutual Influence, Theory and Uses of Animal Magnetism as a Universal Healing Medium, &c.* (ed. Karl Christian Wolfart). Berlin: Nikolai.

Metaphysics [Def. 1]. (n.d.). In OED Online. Oxford University Press. Retrieved 26 June 2016 from http://www.oed.com/view/Entry/117355.

Metaphysics [Def. 2]. (n.d.). In OED Online. Oxford University Press. Retrieved 26 June 2016 from http://www.oed.com/view/Entry/117355.

de la Mettrie, J.O. (1748) Man a Machine. In Thomson, A. (ed.) *Machine man and other writings* (Hardback version transferred to digital print. ed.), 2003, Cambridge: Cambridge University Press.

Michaelis, L. and Menten, M. L. (1913). *Die Kinetik der InvertinwirkungBiochem*. Z.49, 333–369.

Michotte, A.E. (1946/1963). *The Perception of Causality* (Translated by T. R. Miles and E. Miles). London: Methuen.

Michotte, A.E. (1950). À propos de la permanence phénoménale: Faits et théories [On phenomenal permanence: Facts and theories]. *Acta Psychologica*, 7, 298–322.

Michotte, A.E. (1968). The Emotional Significance of Movement. In Arnold, M.B. *The Nature of Emotion*, Harmondsworth, UK: Penguin, pp. 263–78.

Mill, J.S (1874). *Three Essays on Religion.* New York: Henry Holt and Company.

Miller M.B. and Bassler B.L. (2001). Quorum sensing in bacteria. *Annuual Review of Microbiology* 55:165–99.

Miller, S, and Krijnse-Locker, J. (2008). Modification of intracellular membrane structures for virus replication. *Nat Rev Microbiol* 6:363–374.

Miller, S.L. (1953). A production of amino acids under possible primitive Earth conditions. *Science* 117: 528–529.

Millikan, R.G. (1984). *Language Thought and Other Biological Categories*. MIT Press, Cambridge, MA.

Millikan, R.G. (1989a). An ambiguity in the notion of function. *Bio. and Phil.* 4, 172–176.

Millikan, R.G. (1989b). In Defense of Proper Functions. *Philosophy of Science* 56, 288–302.

Millikan, R.G. (1989c). Biosemantics. *The Journal of Philosophy*, Vol. 86, No. 6 (Jun., 1989), pp. 281–297.

Millikan, R.G. (1993). Propensities, Exaptations, and the Brain, in *White Queen Psychology and Other Essays for Alice*, MIT Press, Cambridge, MA, pp. 31–50.

Millikan, R. (1996). On Swampkinds. *Mind & Language*, Vol. 11. No. I March 1996, pp103–177.

Millikan, R. (1999). Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function. *The Journal of Philosophy*, Vol. 96, No. 4 (Apr., 1999), pp. 191-206.

Millikan, R. (2002). Biofunctions: Two paradigms. In Ariew, A., Cummins, R., and Perlman, M. (eds.) *Functions: New essays in the philosophy of psychology and biology*. Oxford University Press: Oxford.

Mills, S.K. and Beatty, J.H. (1979). The Propensity Interpretation of Fitness. *Philosophy of Science*, Vol. 46, No. 2 (Jun., 1979), pp. 263–286.

Milne, A.A. (1927). *Now We Are Six*. New York: Dutton.

Mitchell, R. W. (1996). The history and method of anthropomorphic analysis of anecdotes about animals and God in Western Science. In Mitchell *et al.* (eds.) *Anthropomorphism, Anecdote and Animals*. New York: SUNY Press.

Mitchell, S.D. (1993). Dispositions or etiologies? A comment on Bigelow and Pargetter. *J. Phil.* 90,249–259.

Mitchell, S.D. (1995). Function, Fitness and Disposition. *Biology and Philosophy* 10, 39–54.

Monod, J. (1971). *Chance and Necessity: An essay on the natural philosophy of modern biology*. New York: Alfred A Knopf.

Müller, J. (1833/1840). Elements of Physiology. In Baly, W.M. (transl.) *Handbuch der Physiologie des Menschen für Vorlesungen)*. London: Taylor and Walton, pp. 878–904.

Munson, R. (1972). Biological Adaptation: A Reply. *Philosophy of Science*, 39: 529–32.

Nagarajan, R. and Ganesh, K. (1989). Block copolymer self-assembly in selected solvents, *J. Chem. Phys.* p. 5843.

Nagel, E. (1961). *The Structure of Science.* New York: Harcourt, Brace and World.

Nagel, E. (1977). Teleology revisited: goal-directed processes in biology. *J. Phil.* 74,261-301.

Nagel, T. (2012). *Mind & Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False.* Oxford University Press.

Neander, K. (1983). *Abnormal Psychobiology.* Ph.D. Dissertation, La Trobe University.

Neander, K. (1991a). The teleological notion of "function". Au*str. J. Phil.* 69(4),454–468.

Neander, K. (1991b). Functions as selected effects: the conceptual analyst's defence. *Phil. Sci.* 58,168–184.

Neander, K. (1995). Pruning the Tree of Life, *British Journal for the Philosophy of Science* 46, 59–80.

Needham, J.T. (1750). *Nouvelles observations microscopiques, avec des découvertes Intéressantes sur la composition et la décomposition des corps organizes.* Paris: L.E. Ganeau.

Neumann, J. von (1966). *Theory of Self-Reproducing Automata.* (Edited and completed by Arthur W. Burks. Urbana: University of Illinois Press.

Newton, I. (1687). *Philosophiæ Naturalis Principia Mathematica.* London: Benjamin Motte.

Newton, I. (1692a) Original Letter from Isaac Newton to Richard Bentley, dated 10th December 1692. *The Newton Project.* Retrieved from http://www.newtonproject.ox.ac.uk/ 189.R.4.47, ff. 4A-5, Trinity College Library, Cambridge, UK.

Newton, I. (1692b). Original Letter from Isaac Newton to Richard Bentley, undated, sent in reply to a letter form Mr. Bentley, dated 18th Feburary 1693. *The Newton Project.* Retrieved from http://www.newtonprojet.ox.ac.uk/ 189.R.4.47, ff. 7-8, Trinity College Library, Cambridge, UK.

Newton, I. (1704). *Opticks: or, A Treatise of the Reflexions, Refractions, Inflexions and Colours of Light.* London: Printed for Sam Smith, and Benj. Walford, printers to the Royal Society, at the prince's Arms in St. Paul's Churchyard.

Niazi, M.S. (2017). The Electric Honeycomb; an investigation of the Rose window instability. *R. Soc. open sci.*4:170503.

Nielsen, Julius; Hedeholm, Rasmus B.; Heinemeier, Jan; Bushnell, Peter G.; Christiansen, Jørgen S.; Olsen, Jesper; Ramsey, Christopher Bronk; Brill, Richard W.; Simon, Malene; Steffensen, Kirstine F.; Steffensen, John F. (2016). Eye lens radiocarbon reveals centuries of longevity in the Greenland shark (Somniosus microcephalus). *Science.* 353 (6300): 702–4.

Nietzsche, F. (1887/1897) *Genealogy of Morals: A Polemic.* Ed. Alexander Tille, Trans. William A. Hausemann. London and New York: Macmillan.

Nissen, L. (1993). Four ways of eliminating mind from teleology. *Stud. Hist. Phil. Sci.* 24(1),27–48.

Nissen, L. (1997). *Teleological language in the life sciences*. Lanham.

Novikoff A. B. (1945). The concept of integrative levels and biology. *Science* 101, 209–215.

Novoa, R.R., Calderita, G., Arranz, R., Fontana, J., Granzow, H., and Risco, C. (2005). Virus factories: associations of cell organelles for viral replication and morphogenesis. *Biol Cell* 97:147–172.

Nweeia M.T., Eichmiller, F.C., Hauschka, P.V., Donahue, G.A., Orr, J.R., Ferguson, S.H., Watt, C.A., Mead, J.G., Potter, C.W., Dietz, R., Giuseppetti, A.A., Black, S.R., Trachtenberg, A.J., and Kuo, W.P. (2014). Sensory ability in the narwhal tooth organ system. *Anat Rec* (Hoboken) 297(4): 599–617.

O'Grady, R.T. (1984). Evolutionary Theory and Teleology. *Journal of Philosophy*, 74, 261–301.
Okasha, S. (2002). 'Darwinian Metaphysics: Species and the Question of Essentialism', *Synthese* 131: 191–213.

Oken, L. (1805). *Die Zeugung*. Bamberg: Joseph Anton Goebhardt.

Oken, L. (1831). *Lehrbuch der Naturphilosophie*. Jena: Friedrich Fromann; 2nd edition.

Oparin, A. I. (1964). *Life: Its Nature, Origin and Development*, trans, from the Russian by Ann Synge. New York: Academic Press.

Orgel, L.E. (1968). Evolution of the genetic apparatus. *J Mol Biol.* 1968 Dec; 38(3):381–93.

Orgel, L. E. (1995). Unnatural Selection in Chemical Systems, *Acc. Chem. Res.* 28, 109–118.

Osborn, H.F. (1934). Aristogenesis, the Creative Principle in the Origin of Species. *The American Naturalist* Vol. 68, No. 716.

Oyama, S. (1985). *The Ontogeny of information*, New York, Cambridge University Press.

Oyama, S., Griffiths P.E., & Gray, R.D. (2001). *Cycles of Contingency: Developmental systems and evolution*, Cambridge, Mass., MIT Press.

Paley, W. (1802). *Natural Theology or Evidences of the Existence and Attributes of the Deity*. London: R. Faulder.

Parijs, P. van (1982). *Evolutionary Explanation in the Social Sciences: an emerging paradigm.* Totowa, NJ: Rowman and Littlefield.

Pasteur, L. (1860) in a lecture to the Chemical Society of Paris, "On the Molecular Dissymmetry of Natural Organic Products". Reprinted in The Chemical News and Journal of Industrial Science (3 May 1862), 5, No. 126, p. 248.

Pearse, A.M., Swift, K. (2006). Allograft theory: Transmission of devil facial-tumour disease. *Nature*. 439 (7076): 549.
Pearson, K. (1892). *The Grammar of Science*. London: Walter Scott.

Perez, A.T. (1997). Rose-window instability in low conducting liquids. *Journal of Electrostatics*. vol. 40–41, pp. 141–146.

Perrin, J. (1909). Mouvement brownien et réalité moléculaire [Brownian movement and molecular reality]. Ann. Chim. Phys. 8ième série 18, 5–114.

Persic, M. and Salucci, P. (1992). The baryon content of the Universe. *Monthly Notices of the Royal Astronomical Society*. 258 (1): 14P–18P.

Piaget, J. (1929). *The child's conception of the world*. New York: Harcourt, Brace.

Piaget, J. (1951). Egocentric thought and sociocentric thought. *Sociological studies*, 270-286.

Piaget, J., & Cook, M. T. (1952). *The origins of intelligence in children*. New York, NY: International University Press.

Pigliucci, M. and Boudry, M. (2013). Prove it! The burden of proof game in science vs. pseudoscience disputes. *Philosophia* 42(2):487–502.

Pigliucci, M. and Kaplan, J. (2006). Making Sense of Evolution: The Conceptual Foundations Of Evolutionary Biology. Chicago: University of Chicago Press.

Pinker, S. (1997). Evolutionary psychology: An exchange. *New York Review of Books* 44(15):55–56.

Pittendrigh, C.S. (1958). Adaptation, natural selection and behavior. In Roe, A. and Simpson, G.G. (eds.) *Behavior and Evolution*, Yale University Press, New Haven, CT, pp. 390–419.

Pittendrigh, C.S. (1970). Personal Correspondence with Ernst Mayr, February 26. Pp. 391–392 in Ernst Mayr, *Evolution and the Diversity of Life: Selected Essays*. Cambridge, MA: Belknap Press of Harvard University Press, 1976.

Planck, M. (1901). On the Law of Distribution of Energy in the Normal Spectrum. *Annalen der Physik*, vol. 4, p. 553 ff.

Plantinga, A. (1993). *Warrant and proper function*. Oxford University Press, New York.

Plato, & Jowett, B. (1990). *Charmides, or temperance*. New York: C. Scribner's Sons.

Plato, in Burnet, J. (1924). *Plato's Euthyphro, Apology of Socrates and Crito*.

Plato, & Woodruff, P. (1982). *Hippias major*. Indianapolis: Hackett Pub. Co.

Plato. (1911). *Plato's Phaedo*. Oxford: Clarendon press.

Plato, Anastaplo, G., & Berns, L. (2004). *Plato's Meno*. Newburyport, MA: Focus Pub./R. Pullins Co.

Plato, Grube, G. M. A., & Reeve, C. D. C. (1992). *Republic*. Indianapolis: Hackett Pub. Co.

Plato, Jowett, B., & Pelliccia, H. (1996). *Symposium: The Benjamin Jowett translation*. New York: Modern Library.

Plato, & Jowett, B. (1990). *Theaetetus*. New York: C. Scribner's Sons.

Plato. (1937). *Plato's cosmology; the Timaeus of Plato*. London: New York, Harcourt, Brace, K. Paul, Trench, Trubner & Co. ltd.

Plutarch., Dryden, J., & Clough, A. H. (75AD/2001). *Plutarch's lives* / the Dryden translation, edited with preface by Arthur Hugh Clough ; introduction by James Atlas (Modern Library paperback ed.). New York: Modern Library.

Pohl, R.; Gilman, R.; Miller, G.A.; Pachucki, K. (2013). Muonic hydrogen and the proton radius puzzle. *Annu. Rev. Nucl. Part. Sci.* 63.

Popa, R. (2004). *In Between Necessity and Probability: Searching for the Definition and Origin of Life*. New York: Springer.

Porter, R. and Ogilvie, M., eds. (2000). Wolff, Kaspar Friedrich. *The Biographical Dictionary of Scientists. 3rd ed.*, New York: Oxford University Press.

Povinelli, D.J., Nelson, K.E., and Boysen, S.T. (1990). Inferences about guessing and knowing by chimpanzees (Pan troglodytes). *J Comp Psychol.* 1990 Sep;104(3):203-10.

Povinelli, D.J. and Eddy, T.J. (1996). What chimpanzees know about seeing. *Monographs of the Society for Research in Child Development* 61(2), serial no. 247.

Povinelli, D.J., Vonk, J. (2003) Chimpanzee minds: suspiciously human? *Trends Cogn Sci.* 2003 Apr;7(4):157-160.

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a "theory of mind"? *Behavioral and Brain Sciences*, 4, 515–526.

Preston, B. (1998). Why is a Wing Like a Spoon? A Pluralist Theory of Function. *The Journal of Philosophy*, Vol. 95, No. 5 (May, 1998), pp. 215–254.

Prigogine, I. (1967). Dissipative structures in chemical systems. In *Fast Reactions and Primary Processes in Chemical Kinetics*, Nobel Symposium 5, S.Claesson, ed., Interscience, New York, 371–382.

Prigogine, I., and Lefever, R. (1968). Symmetry Breaking Instabilities in Dissipative Systems II. *The journal of Chemical Physics*, 48, 1695.

Prigogine, I., and Nicolis, G. (1967). Symmetry-Breaking Instabilities in Dissipative Systems. *The journal of Chemical Physics*. 46, 3542.

Prigogine, I., Nicolis, G., Babloyantz, A. (1972). Thermodynamics of Evolution. *Physics Today*: Vol 25, No 12.

Prigogine, I., and Stengers, I. (1984). *Order out of Chaos: Man's new dialogue with nature*. New York: Bantam Books.

Prinz, J. and Clark, A. (2004). Putting Concepts to Work: Some Thoughts for the Twentyfirst Century. *Mind and Language*. Vol. 19, Issue 1, February 2004, pp. 57–69.

Prior, E. (1985). *Dispositions*. Atlantic Highlands, N.J.: Humanities Press.

Pross, A. (2005). Stability in chemistry and biology: Life as a kinetic state of matter. *Pure Appl Chem*, 77:1905–1921.

Pross, A. (2008). How can a chemical system act purposefully? Bridging between life and non-life. *J Phys Org Chem*, 21:724–730.

Pross, A. (2009). Seeking the chemical roots of Darwinism: Bridging between chemistry and biology. *Chem Eur J*, 15:8374–8381.

Purkyně, J.E. (1837). Communication to the Meeting of Naturalists at Prague in 1837. as cited in Müller, J. (1839) *The Intimate Structure of Secreting Glands*. London: Joseph Butler.

Putnam, H. (1962) The Analytic and the Synthetic. In H. Feigl & G. Maxwell (Eds.), Minnesota *Studies in the Philosophy of Science, Volume III*. Minneapolis: University of Minnesota Press.

Putnam, H. (1967). The Nature of Mental States. In *Mind, Language, and Reality: Philosophical Papers*, Vol. 2, 429–440. New York: Cambridge University Press, 1975.

Queller, D.C. & Strassmann, J.E. (2009). Beyond society: the evolution of organismality. *Phil. Trans. R. Soc. B* 2009 364, 3143–3155.

Quine, W.V.O. (1951). Two Dogmas of Empiricism. *The Philosophical Review*, 60, pp. 20–43.

Quine, W.V.O. (1962). The Ways of Paradox. In W.V. Quine (1976) *The Ways of Paradox and Other Essays*, revised and enlarged edition. Cambridge MA: Harvard University Press.

Ramachandran, V.S. and R.L. Gregory (1991). Perceptual filling in of artificially induced scotomas in human vision. *Nature* 350(6320): 699–702.

Ray, J. (1691). The *Wisdom of God Manifested in the Works of Creation*. London: William Innys.

Reichenbach, Karl Baron Von (1850). *Researches on Magnetism, Heat, Light, Electricity, Crystallization, and Chemical Attraction in their relations to vital force.* (Translated and Edited by William Gregory, MD.) London and Edinburgh: Taylor, Walton and Maberly.

Reil, J.C. (1796). Von der Lebenskraft. In *Archiv für die Physiologie*, vol. 1 (1796), pp. 8-162.

Reiss, J.O. (2009). *Not By Design: Retiring Darwin's Watchmaker.* Berkeley CA: University of California Press.

Remak, R. (1852). Über extracellulare Entstehung thierischer Zellen und über die Vermehrung derselben durch Theilung. *Archiv fur Anatomie, Physiologie und wissenschaftiche Medicin* 1852: 47–57.

Resnik D. (1996). Adaptationism: Hypothesis or heuristic? *Biology and Philosophy* 12 (1):39–50.

Ricardo, D. (1817). *On the Principles of Political Economy and Taxation.* London: John Murray.

Richardson, R. (1979). Functionalism and Reductionism. *Philosophy of Science*, 46: 533–558.

Richerson, P. and Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago, University of Chicago Press.

Ritter, W.E. (1919). *The unity of the organism, or the organismal conception of life.* Boston, MA: Gorham Press.

Ritter, W.E. and Bailey, E.W. (1928). *The Organismal Conception: Its Place in Science and Its Bearing on Philosophy*. University of California Press.

Rochat, P., Morgan, R., & Carpenter, M. (1997). Young infants' sensitivity to movement information specifying social causality. *Cognitive Development*, 12, 537–561.

Roels, J.A. (2012). *The Origin and the Evolution of Firms: Information as a Driving Force.* Delft: Delft University Press.

Romey, K. (2018). Ancient Mass Child Sacrifice May Be World's Largest. *National Geographic.* Downloaded from https://news.nationalgeographic.com/2018/04/mass-child-human-animal-sacrifice-peru-chimu-science/

Rosch E.H. (1973). Natural Categories. *Cognitive Psychology* Volume 4, Issue 3, May 1973, Pages 328–350.

Rosenberg, A. (1985). *The Structure of Biological Science.* Cambridge: Cambridge University Press.

Rosenberg, A. (2001a). How Is Biological Explanation Possible? *British Journal for the Philosophy of Science* 52: 735–760.

Rosenberg, A. (2001b). Reductionism in a Historical Science. *Philosophy of Science* 68: 135–163.

Rosenberg, A. (2001c). On multiple realization and special sciences. *Journal of Philosophy*, 98: 365–373.

Rosenberg, A. (2006). *Darwinian Reductionism: Or, How to Stop Worrying and Love Molecular Biology*. Chicago: University of Chicago Press.

Rosenberg, A. (2013). How Jerry Fodor slid down the slippery slope to Anti-Darwinism, and how we can avoid the same fate. *Euro Jnl Phil Sci* 3:1–17.

Rosenberg, A. and Kaplan, D.M. (2005). How to Reconcile Physicalism and Antireductionism about Biology. *Philosophy of Science*: 72, 43–68.

Rosenblueth, A. and Wiener, N. (1950). Purposeful and Non-Purposeful Behavior. *Philosophy of Science*, Vol. 17, No. 4, pp. 318–326.

Rosenblueth, A., Wiener, N., and Bigelow, J (1943). Behavior, Purpose and Teleology. *Philosophy of Science* 10 (1): 21.

Ross, D. (1936). *Aristotle's Physics. A Revised Text with Introduction and Commentary by W.D. Ross.* New York: Clarendon Press.

Rozenblit, L., and F. Keil. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 26:521–562.

Rubin, V.; Thonnard, W. K. Jr.; Ford, N. (1980). Rotational Properties of 21 Sc Galaxies with a Large Range of Luminosities and Radii from NGC 4605 (R = 4kpc) to UGC 2885 (R = 122kpc). *The Astrophysical Journal.* 238:471.

Ruse, M.E. (1971). Reduction, Replacement, and Molecular Biology. *Dialectica*, Volume 25, Issue 1: 39–72.

Ruse, M.E. (1981). Teleology Redux. In Agassi, J. and Cohne R.S. (eds.) *Scientific Philosophy Today: Essays in Honor of Mario Bunge*. New York: Springer.

Ruse, M. (1993). Evolution and progress. *TREE* 8(2),55–59.

Ruse, M., & Travis, J. (Eds.) (2009). *Evolution: The First Four Billion Years.* Cambridge, Massachusetts: The Belknap Press of Harvard University Press

Russell, B. (1912). *The Problems of Philosophy*. London: Henry Holt.

Russell, B. (1918/1985). *The philosophy of logical atomism.* Monist 28: 495–527; 29, 32–63,190–222, 345–380. Reprinted in *The Philosophy of Logical Atomism*, ed. David Pears, 35–155, LaSalle: Open Court, 1985.

Russell, E.S. (1930). *The Interpretation of Development and Heredity: A Study in Biological Method.* Freeport NY: Books for Libraries Press.

Russell, E. S. (1945). *The Directiveness of Organic Activities*. Cambridge: Cambridge University Press.

Ryle, G. (1949). *The Concept of Mind*. Chicago: University of Chicago Press.

Sachs, J. (1995). *Aristotle's Physics: a Guided Study*. New Brunswick NJ: Rutgers University Press.

Sachs, J. (2005). Aristotle: Motion and its Place in Nature. *Internet Encyclopedia of Philosophy*.

Sagan, C. (1979). *Broca's Brain: Reflections on the Romance of Science*. New York: Random House.

Salmon, W.C., (1978). Religion and Science: A New Look at Hume's Dialogues. *Philosophical Studies*, 33, 143–176.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

San Miguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramoza, Z., and Bennetzen, J.L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274, 765–768.

Sarkar G. (1998). *Genetics and Reductionism: A Primer*. Cambridge: Cambridge University Press.

Schaffner, K.F. (1967). Approaches to Reduction. *Philosophy of Science* 34 (2):137–147.

Schaffner, K.F. (1993). *Discovery and Explanation in Biology and Medicine*. Chicago: University of Chicago Press.

Schaller, M. (2003). Ancestral environments and motivated social perception: Goal-like blasts from the evolutionary past. In S. J. Spencer, S. Fein, M. P. Zanna, & J. M. Olson (Eds.), *Motivated social perception: The Ninth Ontario Symposium* (pp. 215–231). Mahwah NJ: Lawrence Erlbaum Associates.

Scheffler, I. (1959). Thoughts on teleology. *The British Journal of Philosophy of Science*, 9, 265–284.

Schleiden, M. J. (1838). *Contributions to Phytogenesi*s. Trans. by H. Smith. Sydenham Soc., London, 1847, p. 231–263.

Schlosser, G. (1998). Self-Re-Production and Functionality: A Systems-Theoretical Approach to Teleological Explanation. *Synthese*, Vol. 116, No. 3 (1998), pp. 303–354.

Schöpenhauer, A. (1883/1969). *The World as Will and Representation*. Trans. E. Payne. New York: Dover.

Schrödinger, E. (1967) *What is Life? The Physical Aspect of the Living Cell*. Cambridge: Cambridge University Press.

Schwann, T. (1839). *Microscopic Investigations on the Accordance in the Structure and Growth of Plants and Animals*. Berlin. (English translation by the Sydenham Society, 1847)

Schwartz, P.H. (2002). The continuing usefulness account of proper functions. In Ariew, A., Cummins, R., and Perlman, M. (eds.) *Functions: New essays in the philosophy of psychology and biology*. Oxford University Press: Oxford.

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3 (3):417–457.

Sellars, W. (1962) Philosophy and the Scientific Image of Man. In Colodny, R. (ed.) *Frontiers of Science and Philosophy*, 1962, Pittsburgh, PA: University of Pittsburgh Press, 35–78.

Shanahan, M. (2005). Cognition, Action Selection, and Inner Rehearsal. In: *Proceedings IJCAI Workshop on Modelling Natural Action Selection*, pp. 92–99.

Shapley, H. (1953). *Climatic Change: Evidence, Causes, and Effects*. Harvard University Press, Cambridge.

Shapiro, A. (2009). William Paley's Lost "Intelligent Design". *History and Philosophy of the Life Sciences - Vol. 31*, no. 1.

Simmons, A. (2001). Sensible Ends: Latent Teleology in Descartes' Account of Sensation. *Journal of the History of Philosophy* 39.1. pp. 49–75.

Simon, H. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6), 467–482.

Simpson, G.G. (1949) *The Meaning of Evolution: A Study of the History of Life and of Its Significance for Man*. New Haven: Yale University Press.

Simpson, G.G. (1953a). *The Major Features of Evolution*. New York: Columbia University Press.

Simpson, G.G. (1953b). *Life of the Past: An Introduction to Paleontology*. New Haven CT: Yale University Press.

Simpson, G.G. (1964). *This View of Life: The world of an evolutionist*. New York: Harcourt, Brace & World.

Skinner, B. F. (1974). *About Behaviorism*. New York: Vintage Books.

Smith, E.E., and Medin, D.L. (1981). *Categories and Concepts*. Cambridge MA: Harvard University Press.

Smullyan, R. (1961). *Theory of Formal Systems*. Princeton, N.J.: Princeton University Press.

Smuts, J.C. (1926). *Holism and Evolution*. London: MacMillan and Co.

Sober, E. (1984a). *The Nature of Selection*. MIT Press, Cambridge, MA.

Sober, E. (ed.) (1984b). *Conceptual Issues in Evolutionary Biology*. MIT Press, Cambridge, MA.

Sober, E. (1993). *Philosophy of biology*. Boulder, CO: Westview Press.

Sojo, V., Herschy, B., Whicher, A., Camprubí, E., and Lane, L. (2016). The Origin of Life in Alkaline Hydrothermal Vents. *Astrobiology*. 16: 181–197.

Sommerhoff, G. (1950). *Analytical Biology*. New York: Oxford University Press.

Sommerhoff, G. (1969). The abstract characteristics of living systems. In F. E. Emery (Ed.), *Systems thinking* (pp. 147–202). Middlesex: Penguin.

Sommerhoff, G. (1974). *Logic of the living brain*. New York: Wiley.

Sorabji, R. (1964). Function. *The Philosophical Quarterly* 14: 289–302.

Spencer, H. (1864). *The Principles of Biology*. London: Williams and Norgate.

Sperry, R.W. (1965). Mind, Brain, and Humanist Values. In John R. Platt (ed.), *New Views on the Nature of Man*. Chicago: University of Chicago Press.

Sperry, R. W. (1969). A modified concept of consciousness. *Psychol. Rev.* 76, 532–536.

Sperry, R.W. (1980). Mind-Brain Interaction: Mentalism, Yes; Dualism, No. *Neuroscience* Vol. 5. pp. 195–206.

Spinoza, B. (1677/1996). *Ethics*. London: Penguin.

Stevenson, K.B., Harrington, J., Lust, N.B., Lewis, N.K., Montagnier, G., Moses, J.I., Visscher, C., Blecic, J., Hardy, R.A., Cubillos, P., Campo, C.J. (2012). Two nearby sub-Earth-sized exoplanet candidates in the GJ 436 system. *The Astrophysical Journal*. Vol 755, Issue 1, Article 9.

Stich, S. (1993). Moral Philosophy and Mental Representation. In M. Hechter, L. Nadel & R. Michod (eds.), *The Origin of Values*. New York: Aldine de Gruyer.

Strughold, H. (1953). *The Green and Red Planet: A Physiological Study of the Possibility of Life on Mars*. University of New Mexico Press.

Styer, D. F. (2000). Insight into entropy. *Am. J. Phys.* 68, 1090-1096.

Sushkov, A. O.; Kim, W. J.; Dalvit, D. A. R.; Lamoreaux, S. K. (2011). New Experimental Limits on Non-Newtonian Forces in the Micrometer Range. *Physical Review Letters* 107 (17): 171101.

Sutton, W.S. (1902). On the morphology of the chromosome group in Brachystola magna. *Biol. Bull.* 4: 24–39.

Sutton, W.S. (1903). The chromosomes in heredity. *Biol. Bull.* 4: 231–251. Partial reproduction in: Classic Papers in Genetics (1959) (Peters, J.A., ed.). Prentice-Hall, Englewood Cliffs, pp. 27–41.

Swetz, F., Fauvel, J., Bekken, O., Johansson, B., Katz, V. (Eds.) (1995). *Learn From the Masters*. The Mathematical Association of America.

Symons, D. (1981). The semantics of ultimate causation. *The Evolution of Human Sexuality*. Oxford: Oxford University Press.

Taleb, N. (2001). *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. New York: Random House and Penguin.

Taleb, N. (2007). *The Black Swan: The Impact of the Highly Improbable*. New York: Random House and Penguin.

Taylor, C.C.W. (1999). *The Atomists: Leucippus and Democritus: Fragments*. Toronto: University of Toronto Press.

Taylor, P.W. (1986). *Respect for Nature: A Theory of Environmental Ethics*. Princeton: Princeton University Press.

Taylor, R. (1950a). Comments on a mechanistic conception of purposefulness. *Philosophy of Science*. 17 (4), 310–317.

Taylor, R. (1950b). Purposeful and non-purposeful behavior: A rejoinder. *Philosophy of Science*. 17 (4), 327–332.

Tegmark, M. (2014). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. New York: Alfred A. Knopf.

Teilhard de Chardin, P. (1955). *Le Phénomène Humaine*. New York: Harper.

Thomas, D. (1952/1955) A Child's Christmas in Wales. Norfolk: New Directions.

Thomas, E.M., (1993). *The Hidden Life of Dogs*. Boston: Houghton Mifflin.

Thomson, W. (Lord Kelvin) (1852). On a Universal Tendency in Nature to the Dissipation of Mechanical Energy. *Proceedings of the Royal Society of Edinburgh III*, April 19, 1852, p. 139–142.

Thompson, E. (2007). *Mind in Life : Biology, Phenomenology, and the Sciences of Mind.* Cambridge MA: Harvard University Press.

Thompson, N.S. (1987). The Misappropriation of Teleonomy. In, P.P.G.B. Bateson and P. Klopfer (eds.) *Persp. Ethol.* 7,259–274.

Tonomura, A., *et al.* (1986). Evidence for Aharonov-Bohm Effect with Magnetic Field Completely Shielded from Electron Wave, *Phys. Rev. Lett.* 56, 792.

Tooby, J. and L. Cosmides (1992). The Psychological Foundations of Culture. In H. Barkow, L. Cosmides and J. Tooby (eds.), *The Adapted Mind*, New York: Oxford University Press, pp. 19–136.

Treviranus, G.R. (1831,1832). *Die Erscheinungen und Gesetze des organischen Lebens.* Bremen.

Trifonov, E. (2011). Vocabulary of Definitions of Life Suggests a Definition. *Journal of Biomolecular Structure & Dynamics,* Volume 29, Issue Number 2.

Trimble, V. (1987). Existence and nature of dark matter in the universe. *Annual Review of Astronomy and Astrophysics*. 25: 425–472.

Tsang, M. and Caves, C. (2012). Evading Quantum Mechanics: Engineering a Classical Subsystem within a Quantum Environment. *Phys. Rev. X* 2, 031016.

Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind* 49: 433-460.

Tylor, E. (1871). *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art, and Custom*. London: J. Murray.

Vallery-Radot, René. (1928). *The Life of Pasteur*. R.L. Devonshire.

Varela, F. G., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Biosystems*, 5(4):187–196.

Virchow, R. (1858). *Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre.* Berlin: A. Hirschwald.

Voltaire (1901) [1734]. "On the Existence of God". *The Works of Voltaire: The Henriade: Letters and miscellanies XXI*. trans. William F. Fleming. Werner. pp. 239–240.

Vygotsky, L. S. (1986). *Thought and language.* Cambridge, MA: MIT Press.

Wade, L. (2013). "Llullaillaco Maiden" May Have Been Drugged Before Sacrificed. *Science Magazine*, 29 July 2013. Retreived from http://news.sciencemag.org/2013/07/llullaillaco-maiden-may-have-been-drugged-sacrificed.

Wald, A. (1943). *A Method of Estimating Plane Vulnerability Based on Damage of Survivors*. Statistical Research Group, Columbia University.

Wallace, A.R. (1858). On the Tendency of Varieties to depart indefinitely from the Original Type. In Darwin, C. R. and A. R. Wallace. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. [Read 1 July] *Journal of the Proceedings of the Linnean Society of London. Zoology* 3 (20 August): 45–50.

Walsh, D.M. (1996). Fitness and Function. *The British Journal for the Philosophy of Science*, Vol. 47, No. 4 (Dec., 1996), pp. 553–574.

Walsh, D.M. (2006). Organisms as natural purposes: the contemporary evolutionary perspective. *Stud Hist Philos Biol Biomed Sci*. 37(4):771–91.

Walsh, D.M. and Ariew, A. (1996). A Taxonomy of Functions. *Canadian Journal of Philosophy*, Vol. 26, No. 4, pp. 493–514.

Walsh, K. J. (2009). Asteroids: When planets migrate. *Nature* 457, 1091–1093.

Walsh, K. J., Morbidelli, A., Raymond S. N., O'Brien, D. P., and Mandell A. M. (2011). A low mass for Mars from Jupiter's early gas-driven migration. *Nature* 475, 206–209.

Watson, J.B. (1930). *Behaviorism*, W. W. Norton and company, New York.

Watson J.D. and Crick, F.H.C. (1953). A Structure for Deoxyribose Nucleic Acid. *Nature* 171, 737–738.

Weber, M. (2005). *Philosophy of Experimental Biology*. Cambridge MA: Cambridge University Press.

Weismann, A. (1909). The Selection Theory. In A.C. Seward (ed.) *Darwin and Modern Science*. Cambridge: Cambridge University Press, 23–86.

Wertheimer, M. (1923). Laws of Organization in Perceptual Forms. First published as Untersuchungen zur Lehre von der Gestalt II, in Psycologische Forschung, 4, 301–350. Translation published in Ellis, W. (1938). A source book of Gestalt psychology (pp. 71–88). London: Routledge & Kegan Paul.

West-Eberhard, M. J. (1992). Adaptation: current usages. In E.F. Keller and E.A. Lloyd (eds.) *Keywords in Evolutionary Biology*, Harvard University Press, Cambridge, MA, pp. 13–18.

Wiener, N. (1961). *Cybernetics or Control and Communication in the Animal and the Machine*, 2nd edition, Cambridge, MA: MIT Press.

Wigner, E. P. (1960). The Unreasonable Effectiveness of Mathematics in the Natural Sciences. In *Communications in Pure and Applied Mathematics*, vol. 13, No. I (February 1960). New York: John Wiley & Sons.

Williams G.C. (1957). Pleiotropy, natural selection, and the evolution of senescence. *Evolution* 11: 398–411.

Williams, G.C. (1966). *Adaptation and Natural Selection,* Princeton, Princeton University Press.

Williams, G.C. (1992). *Natural Selection: domains, levels, and challenges*. Oxford University Press, New York.

Williams, M.B. (1970). Deducing the consequences of evolution: A mathematical model. *Journal of Theoretical Biology*, 29:343–385.

Wilson, D.S. and Sober, E. (1994). Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences*. 17 (4): 585–654.

Wilson, A.S., Brown, E.L., Villa, C., Lynnerup, N., Healey, A., Ceruit, M.C., Reinhard, J., Previgliano, C.H. Araoz, F.A., Diez, J.G., and Taylor, T. (2013). Archaeological, radiological, and biological evidence offer insight into Inca child sacrifice. *PNAS* vol. 110, no. 33.

Wilson, E.B. (1925). *The Cell in Development and Heredity*, 3rd edition. Macmillan, New York.

Wilson, E.O. (1971). *The Insect Societies.* Cambridge MA: Harvard University Press.

Wilson, E. O. (1975). *Sociobiology*, Harvard University Press, Cambridge, Mass.

Wilson, E.O. (2009). *The Superorganism: The Beauty, Elegance, and Strangeness of Insect Societies.* W.W. Norton and Company.

Wimsatt, W.C. (1972). Teleology and the Logical Structure of Function Statements. *Stud. Hist. Phil. Sci.*, 3, no. 1.

Wimsatt, W. C. (1986). Forms of aggregativity. In M. G. Grene, A. Donagan, A. N. Perovich, & M. V. Wedin (eds.), *Human Nature and Natural Knowledge* (pp. 259–291). Dordrecht: Reidel.

Wimsatt, W. (2002). Functional organization, analogy, and inference. In Ariew, A., Cummins, R., and Perlman, M. (eds.) *Functions: New essays in the philosophy of psychology and biology*. Oxford University Press: Oxford.

Wittgenstein, L. (1953/2001). *Philosophical Investigations*. Blackwell Publishing.

Wöhler, F. (1828). "Über künstliche Bildung des Harnstoffs". *Annalen der Physik und Chemie* 88 (2): 253–256.

Woese, C. (1967). *The Genetic Code: The Molecular Basis for Genetic Expression.* Harper and Row, New York

Wolff, C.F. (1759). *Theoria Generationis*. Doctoral Dissertation, University of Halle, Germany.

Woodfield, A. (1976). *Teleology.* London: Cambridge University Press.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69, 1–34.

Wouters, A. (2005). The Functional Perspective of Organismal Biology. In T.A.C. Reydon and L. Hemerik.,(eds.), *Current Themes in Theoretical Biology*, 33–69.

Wright, L. (1968). The Case against Teleological Reductionism. *The British Journal for the Philosophy of Science*, Vol. 19, 1968, 211–23.

Wright, L. (1973/1984). Functions. *Phil. Rev.* 82,139-168. Reprinted in Sober (1984b), pp. 347–368.

Wright, L. (1976). *Teleological Explanations*, University of California Press, Berkeley.

Wynne-Edwards, V.C. (1962). *Animal Dispersion in Relation to Social Behavior.* Oliver & Boyd, London.

Yang, Y., Tang, J., Wu, W.-M., Zhao, J., Song, Y., Gao, L., Yang, R., and Jiang, L. (2015). Biodegradation and Mineralization of Polystyrene by Plastic-Eating Mealworms: Part 1. Chemical and Physical Characterization and Isotopic Tests. *Environmental Science and Technology*. 49 (20): 12080–12086.

Yeats, W.B. (1921). The Second Coming. In *Michael Robartes and the Dancer*. Dundrum, Ireland: Cuala Press.

Yonge, C.D. (Trans.) (1853) The Lives and Opinions of Eminent Philosophers by Diogenes Laertius. London: J. Haddon and Son.

Young, T. (1804). The Bakerian Lecture: Experiments and Calculations Relative to Physical Optics. *Philosophical Transactions of the Royal Society of London, 94*, 1–16.

Zahid, A. (2004). The Vermiform Appendix: Not a Useless Organ. *J. Coll Physicians Surg. Pak.* 14(4):256–8.

Zhabotinsky, A.M. (1964). Периодический процесс окисления малоновой кислоты растворе [Periodical process of oxidation of malonic acid solution]. Биофизика. 9: 306–311.

Zhu J., Miller M.B., Vance R.E., Dziejman M., Bassler B.L., and Mekalanos J.J. (2002). Quorum-sensing regulators control virulence gene expression in *Vibrio cholerae*. *Proc Natl Acad Sci* 99(5):3129–34.

Curriculum Vitae

Matthew M. Hurley
matthew.m.hurley@gmail.com

Education

Indiana University—Bloomington, IN — 2008 – 2018
Ph.D. in Cognitive Science
Committee Chair: Dr. Douglas R. Hofstadter
Committee: Dr. Colin Allen, Dr. Randall Beer, and Dr. Hamid Ekbia

Tufts University—Medford, MA — 2004 – 2006
B.S. in Computer Science and B.S. in Cognitive Science
Committee Chair: Dr. Daniel C. Dennett
Committee: Dr. Reginald B. Adams, Jr., and Dr. Barry Trimmer

Commercial and Research Experience

Big Company—Andover, MA — 2004 – 2008
Independent Contractor, Software Engineer, Database Architect

Failed Dotcom-Era Startup—Waltham, MA — 2001 – 2004
Software Engineer, Database Engineer

Independent Researcher—Berlin, Germany — 1999 – 2000
Data Mining and Market Dynamics Analyst

Hurley Construction, Inc.—Lexington, MA — 1998 – 1999
Home Renovation and Remodeling Contractor

Independent Explorer—The World — 1996 – 1999
Vagabond, Bandit

Local Restaurants—Reading, MA — 1993 – 1995
Pizza Maker, Delivery Boy

Various Local Newspapers—Reading, MA — 1985 – 1995
Paperboy

Published Books

Inside Jokes: Using Humor to Reverse-Engineer the Mind 2011
Co-authored with Daniel C. Dennett and Reginald B. Adams, Jr.
Cambridge MA: MIT Press.

Other Publications

Hurley, M.M., Dennett, D.C., & Adams Jr., R.B. (2013, August 8) "The Complicated Search for the Origins of Humor." *Huffington Post*. http://www.huffingtonpost.com

Hurley, M.M., Dennett, D.C., & Adams Jr., R.B. (2011) "Q&A with James Garvey." *The Philosopher's Magazine*. Issue 53, 2nd Quarter, 2011, p114-115.